

UNIVERSIDADE DE LISBOA  
FACULDADE DE CIÊNCIAS  
DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA



## **Segmentação de geoinformação: Um caso de estudo na área da Diabetes**

Rafael Marques dos Santos

**Mestrado em Sistemas de Informação Geográfica – Tecnologias e Aplicações**

Trabalho de Projeto orientado por:  
Cristina Catita  
Rita Matildes

# Agradecimentos

Em primeiro lugar gostaria de agradecer às duas pessoas que graças a elas estou aqui a escrever estas palavras, a minha mãe e o meu irmão. Obrigado pelo apoio constante, compreensão e ajuda ao longo de todo o meu percurso académico, que sem dúvida nenhuma são parte da razão por ter conseguido alcançar e finalizar este processo.

Em segundo lugar gostaria de agradecer a outras duas pessoas que foram dois pilares deste projeto, às minhas orientadoras Cristina Catita e Rita Matildes. Primeiramente, gostaria de agradecer à professora Cristina pela total disponibilidade e apoio durante todo este processo letivo no mestrado, desde o início ao fim, incluindo neste projeto, e com certeza será alguém que terei sempre na memória como exemplo de como devemos desempenhar a nossa função profissional e sermos empáticos com o próximo. À Rita Matildes, o meu obrigado pela orientação neste projeto, pela simpatia e preocupação, mesmo quando não era preciso dizer qualquer palavra.

Não poderia deixar de agradecer à empresa ESRI Portugal por me permitir integrar na empresa, embora por breves momentos, e me fornecer toda a tecnologia necessária e um lugar nas instalações para poder efetuar lá este projeto. No entanto, gostaria de deixar um particular agradecimento à senhora Fátima Silva pela gentileza e interesse demonstrado neste projeto.

Gostaria de destacar os meus dois colegas José Alves e Sérgio Rosa, que ao longo deste mestrado fomo-nos habituando à companhia e principalmente ao debate de ideias e companheirismo.

Gostaria de agradecer à Bárbara Simões e ao Miguel Santos pelo constante apoio e presença na minha vida e o quão importantes são para mim, e aos restantes amigos por todo o apoio e por me ajudarem a descontraír em tantos momentos

Para finalizar, a todas as pessoas que de forma direta ou indireta me apoiaram e tornaram possível a realização deste projeto, o meu obrigado.

# Resumo

A tecnologia SIG é um instrumento prático para o apoio nos cuidados de saúde uma vez que graças a esta podemos compreender desigualdades num determinado território, ajudar em tomadas de decisão e na gestão de informação. Através destes é possível verificar padrões espaciais de doenças, o que permite elaborar medidas de compreensão e mitigação.

Neste presente projeto pretendeu-se criar uma segmentação da população de Portugal, a nível de freguesias, de forma a entender quais as características que influenciam um melhor comportamento da população diabética relativamente à sua doença. Utilizou-se o algoritmo “K-means” para criar esta segmentação e com a ajuda do índice Local de Moran foi feita uma análise espacial dos atributos e dos grupos de forma a entender se existiria também algum padrão espacial.

Com o desenvolvimento deste projeto foi possível verificar que em Portugal existe um padrão específico numa zona interior Centro-Norte, sendo composto por uma população mais idosa e com menor nível de rendimentos e escolaridade, em que estes indivíduos apresentam um menor controlo na sua doença uma vez que eram aqueles que apresentam um menor valor de proporção de realização de exames ou de valores saudáveis nestes mesmos. Freguesias onde existiam indivíduos com um melhor nível de rendimento e uma faixa etária mais reduzida apresentam melhores indícios de comportamentos positivos sobre a sua doença, uma vez que estas proporções tinham valores mais elevados comparativamente às restantes freguesias.

Foram criados quatro produtos de informação, nomeadamente três Operational Dashboards distintos, que demonstram o resultado da segmentação e dados da diabetes, performance dos agrupamentos de centro de saúde (ACES) e áreas regionais de saúde (ARS) e todos os atributos públicos disponíveis a nível de Municípios de forma a demonstrar o estado destes a nível nacional. Foi criado também um visualizador com as áreas de influência dos hospitais públicos de Lisboa de forma a demonstrar quais são as freguesias que são inclusas em horas de trânsito e fora destas.

Palavras-Chave: SIG, Segmentação, Moran, Freguesia, População

# Abstract

GIS technology is a practical tool to support health care, as it helps to understand inequalities in a given territory, to decision making and information management. Thanks to it, it is possible to check disease spatial patterns, which allows to take measures of mitigation.

In this project it was intended to create a segmentation of the population of Portugal, at parish level, to understand which characteristics influences a batter behavior of the diabetic population towards their disease. The “K-means” algorithm was used to create this segmentation, and, with the help of Local Moran Index, a spatial analysis of the attributes and groups was made to understand it where would exist any spatial pattern.

With the development of this project it was possible to verify that in Portugal there is a specific pattern in a country-side North-Central zone, being composed of an older population with lower level of income and education, in which these individuals have less control in their disease since they were those that presented a lower value of proportion of examinations or healthy values in these same ones. The parishes that had individuals with a better level of income and a lower age group, presented a better evidence of positive behaviors on their disease, since these proportions had higher values compared to the other parishes.

Four information products were created, namely three distinct Operational Dashboards, which demonstrate the results of the segmentation and Diabetes data, the performance of Agrupamentos de Centro de Saúde (ACES) and Regional Health Areas (ARS) and one that shows all the public attributes available at municipality level, to demonstrate the state of these at national level. It was also created a Visualizer with the areas of influence of the public hospitals of Lisbon to demonstrate which are the parishes that are included in those in hours of traffic and outside of these.

Keywords: GIS, Segmentation, Moran, Parish, Population



# Índice

1.	Introdução .....	1
1.1	Enquadramento do Projeto .....	1
1.2	Objetivos .....	4
1.3	Estrutura do Relatório.....	5
2.	Fundamentos Teóricos e Estado de Arte .....	6
2.1	SIG na área da Saúde.....	6
2.1.1	Monitorização de Doenças .....	6
2.1.2	Análise de Risco .....	8
2.1.3	Planeamento e Acesso a Cuidados de Saúde .....	11
2.1.4	Construção do Perfil de Saúde da Comunidade .....	11
2.2	Diabetes.....	11
2.3	Diabetes e SIG.....	13
2.4	Segmentação na área da saúde.....	16
2.5	Análise de Autocorrelação Espacial Local .....	22
3.	Segmentação .....	25
3.1	Metadados .....	25
3.2	Metodologia .....	25
3.2.1	Seleção das variáveis relevantes .....	26
3.2.2	Modelo Utilizado .....	26
3.2.3	Autocorrelação Espacial Local .....	30
3.3	Análise da Segmentação.....	31
3.3.1	Análise Exploratória dos Atributos.....	33
3.3.2	Análise Espacial de Clusters e respetiva caracterização .....	45
3.4	Análise Exploratória de dados de Diabetes.....	57
4.	Produtos de Informação.....	67
4.1	Metadados .....	67
4.2	Metodologia .....	69
4.2.1	<i>Operations Dashboard</i> de monitorização dos cuidados de saúde a nível de ARS e ACES .....	69
4.2.2	<i>Operations Dashboard</i> de monitorização/gestão proativa dos cuidados de saúde a nível de Municípios e Distritos .....	71
4.2.3	<i>Operations Dashboard</i> com o resultado da segmentação e dados da Diabetes .....	72
4.2.4	<i>Visualizador</i> com áreas de influência de Hospitais em Lisboa .....	73
4.3	Resultados .....	76
5.	Conclusões Finais e Trabalhos Futuros .....	78
	Referências Bibliográficas .....	80

Anexos .....	83
Anexo I – Mapas com a Distribuição Espacial dos Atributos Utilizados na Segmentação .....	83
Anexo II – Dados relativos à Diabetes .....	137
Anexo III – Resultados da Segmentação com 2 e 3 Clusters .....	153
Anexo IV – Dados utilizados nos Produtos de Informação .....	155
Anexo V – Dados utilizados no Dashboard dos ACES .....	157
Anexo VI - Código Python.....	159

# Lista de Figuras

Figura 2.1 - SIG e algumas variáveis.....	6
Figura 2.2 - Distribuição das Unidades de Tratamento de Hanseníase (UTH) e a densidade de casos desta doença para o período dos dados. (Extraído de [15]) .....	7
Figura 2.3 - Localização das ocorrências de hanseníase e áreas de ocupação subnormais em Bayeux. (Extraído de [15]) .....	8
Figura 2.4 - Principais fontes estacionárias de poluição e população minoritária em Bronx. (Adaptado de [16]).....	9
Figura 2.5 - Hospitalização devido a Asma. (Adaptado de [16]).....	9
Figura 2.6 - Buffers de Proximidade de Fontes de Poluição. (Adaptado de [16]).....	10
Figura 2.7 - Insulina a atuar num organismo saudável. (Extraído de [18].).....	12
Figura 2.8 - Insulina a atuar num organismo portador de Diabetes tipo 1. (Extraído de [18]).....	12
Figura 2.9 – Insulina a atuar num organismo portador de Diabetes tipo 2. (Extraído de [18]) .....	13
Figura 2.10 - Número de Incidências de Diabetes na Suécia. (Adaptado de [21]).....	14
Figura 2.11 - Taxa de Incidência de Diabetes nos 5 condados da zona sudeste da Suécia. (Extraído de [21]) .....	14
Figura 2.12 - Identificação de populações alvo para intervenção na área da Diabetes em Michigan. (Adaptado [23]).....	15
Figura 2.13 - Índice de Privação para o Risco de Diabetes na área metropolitana do Porto (Extraído de [24]) .....	16
Figura 2.14 - Pseudocódigo do Algoritmo K-means ( Adaptado de [30]) .....	18
Figura 2.15 - Primeiro Passo do algoritmo "K-means" (Adaptado de [27]).....	18
Figura 2.16 - Segundo e Terceiro passo do algoritmo "K-means"(Adaptado de [27]).....	19
Figura 2.17 - Segmentos do projeto Healthy Foundations. (Extraído de[25]) .....	20
Figura 2.18 - Grupo "Hedonistic Immortals" - Healthy Foundations. (Extraído de [25]) .....	20
Figura 2.19 - Grupo "Live for Todays" - Healthy Foundations. (Extraído de [25]).....	20
Figura 2.20 - Grupo "Unconfident Fatalists" - Healthy Foundations. (Extraído de [25]).....	21
Figura 2.21 - Grupo "Health-Conscious Realists" - Healthy Foundations. (Extraído de [25]).....	21
Figura 2.22 - Grupo "Balanced Compensators" - Healthy Foudations. (Extraído de [25]) .....	21
Figura 2.23 - Estimativas baseadas em modelo de Taxa de Natalidade na Adolescência em 2003 e 2012 e respetivos clusters e outliers espaciais (Adaptado de [37]).....	23
Figura 2.24 - Clusters de valores elevados e reduzidos de índice de massa corporal (Extraído de [38]) .....	24
Figura 3.1 - Três soluções de Segmentação geradas com 2, 3 e 4 clusters .....	29
Figura 3.2 - Exemplo do Output do Índice Local de Moran .....	31
Figura 3.3 - Mapa com o resultado da Segmentação de Geoinformação em Portugal .....	32
Figura 3.4 - Mapas dos atributos referentes ao Género (à esquerda género Feminino, à direita Masculino) e respetivos índices locais de Moran .....	33
Figura 3.5 - Mapas dos atributos referentes a Faixas etárias de Jovens de Adolescentes (da esquerda para a direita: 0 a 4, 5 a 9, 10 a 13 e 14 a 19) e respetivos índices locais de Moran .....	35
Figura 3.6 - Mapas dos atributos referentes a Faixas Etárias de Adultos e Idosos (da esquerda para a direita: 20 a 24, 25 a 65 e mais de 65 anos) e respetivos índices locais de Moran .....	35
Figura 3.7 - Mapas dos atributos referentes ao Ensino (da esquerda para a direita: 1º, 2º e 3º ciclo) e respetivos índices locais de Moran.....	37
Figura 3.8 - Mapas dos atributos referentes ao Ensino ((da esquerda para a direita: Secundário, Pós-Secundário e Superior) e respetivos índices locais de Moran .....	38

Figura 3.9 - Mapas dos atributos referentes à Empregabilidade ((da esquerda para a direita: Desempregados à procura de 1º emprego, Desempregados à procura de emprego e Empregados e respetivos índices locais de Moran.....	39
Figura 3.10 - Mapas dos atributos referentes à Empregabilidade (da esquerda para a direita: Pensionistas e Reformados e Sem Atividade Económica) e respetivos índices locais de Moran .....	40
Figura 3.11 - Mapas dos atributos referentes ao Rendimento (da esquerda para a direita: 1º, 2º e 3º Quintil) e respetivos índices locais de Moran.....	41
Figura 3.12 - Mapas dos atributos referentes ao Rendimento tendo em conta famílias pertencentes ao 4º e 5º Quintil e respetivos índices locais de Moran .....	42
Figura 3.13 - Mapas da Densidade Populacional e respetivo índice local de Moran .....	43
Figura 3.14 - Mapas do Gasto em Álcool e tabaco e respetivos índices locais de Moran .....	44
Figura 3.15 - Mapa com a localização das freguesias pertencentes ao Cluster 1 .....	45
Figura 3.16 - Mapa com exemplo de Outlier do Cluster 1 e a análise com índice local de Moran do atributo “65 ou mais anos de idade” e “Faixa Etária 25 a 64 anos”.....	48
Figura 3.17 - Mapa com a localização das freguesias pertencentes ao Cluster 2 .....	49
Figura 3.18 - Mapa com exemplo de Outlier do Cluster 2 e a análise com índice local de Moran .....	51
Figura 3.19 - Mapa com a localização das freguesias pertencentes ao Cluster 3 .....	52
Figura 3.20 - Mapa com exemplo de Outlier do Cluster 3 e a análise com índice local de Moran .....	54
Figura 3.21 - Mapa com a localização das freguesias pertencentes ao Cluster 4 .....	55
Figura 3.22 - Mapa com exemplo de Outlier do Cluster 4 e a análise com índice local de Moran .....	57
Figura 3.23 – Mapa com os Agrupamentos de Centros de Saúde (ACES) em Portugal .....	58
Figura 3.24 - Série temporal de Utentes inscritos com Diabetes com exame dos pés realizado (2014, 2015, 2016 e 2017).....	60
Figura 3.25 - Série temporal de Proporção de Utentes inscritos com Diabetes com exame dos pés realizado (2014, 2015, 2016 e 2017) .....	61
Figura 3.26 - ACES Lisboa Central com resultados da Segmentação e índices locais de Moran a alguns atributos .....	62
Figura 3.27 - Série temporal de Utentes inscritos com Diabetes com último resultado de HgbA1c inferior ou igual a 8,0% (2014, 2015, 2016 e 2017) .....	63
Figura 3.28 - Série temporal de Proporção de Utentes inscritos com Diabetes com último resultado de HgbA1c inferior ou igual a 8,0% (2014, 2015, 2016 e 2017).....	64
Figura 3.29 - ACES Cavado I com resultados da Segmentação e índices locais de Moran a alguns atributos.....	65
Figura 3.30 - Localização de ACES e índice local de Moran dos atributos "Gasto em Álcool" e "Gasto em Tabaco" .....	65
Figura 4.1 - Exemplo da localização dos dados com ACES como resolução espacial (Mapa de Acesso a Consultas Médicas pela População Inscrita) e Exemplo da localização dos dados com Instituições Hospitalares como resolução espacial (Mapa de Atendimentos em Urgência Hospitalar por Triagem de Manchester).....	68
Figura 4.2 - Mapas respeitantes à localização de Hospitais Públicos, Centros de Saúde e Farmácias e Parafarmácias no concelho de Lisboa.....	68
Figura 4.3 – Mapa com as Áreas Regionais de Saúde (ARS) em Portugal.....	70
Figura 4.4 - Workflow da adição dos dados para a feature ACES/ARS .....	70
Figura 4.5 - Mapas com os Municípios (esquerda) e Distritos (direita) de Portugal.....	72
Figura 4.6 - Mapa com Áreas de Influência referente ao Hospital Egas Moniz na hora de trânsito (esquerda) e ao fim de semana (direita).....	74
Figura 4.7 - Modelo de geoprocessamento .....	75

Figura 4.8 - Mapa com as Freguesias na área de influência do hospital Egas Moniz (Hora de Trânsito)	75
Figura 4.9 - Operational Dashboards de monitorização dos cuidados de saúde a nível de ARS e ACES	76
Figura 4.10 - Operational Dashboard de monitorização/gestão proativa dos cuidados de saúde a nível de Municípios	77
Figura 4.11 - Operational Dashboards que demonstra o resultado da segmentação e contém dados da Diabetes	77
Figura 4.12 - Visualizador com áreas de influência de Hospitais em Lisboa	77

## Lista de Gráficos

Gráfico 1.1 - Taxa de Prevalência de Diabetes por Faixa Etária e Género. (Extraído de [9]).	2
Gráfico 1.2 - Número de novos casos de Diabetes por 10 000 indivíduos por ano. (Extraído de [9])	2
Gráfico 1.3 - Número de Óbitos devido a Diabetes em Portugal por ano. (Extraído de [11])	3
Gráfico 1.4 - Número total de óbitos em Portugal por ano. (Extraído de [11])	3
Gráfico 3.1 - Estatística Pseudo-F do número de Clusters	28
Gráfico 3.2 - Número de freguesias por cluster	31
Gráfico 3.3 - Atributos referentes ao Género por Cluster	34
Gráfico 3.4 - Atributos referentes ao Faixa etária de Jovens e Adolescentes por cluster	36
Gráfico 3.5 - Atributos referentes a Faixas Etárias de Adultos e Idosos por cluster	36
Gráfico 3.6 - Atributos referentes ao Ensino por cluster	38
Gráfico 3.7 - Atributos referentes à Empregabilidade por cluster	40
Gráfico 3.8 - Atributos referentes ao Rendimento por cluster	42
Gráfico 3.9 - Densidade Populacional por cluster	43
Gráfico 3.10 - Gasto em Álcool e Tabaco por cluster	45
Gráfico 3.11 - Faixa Etária (Cluster 1)	46
Gráfico 3.12 - Empregabilidade (Cluster 1)	46
Gráfico 3.13 - Ensino (Cluster 1)	47
Gráfico 3.14 - Rendimento (Cluster 1)	47
Gráfico 3.15 - Faixa etária (Cluster 2)	49
Gráfico 3.16 - Empregabilidade (Cluster 2)	50
Gráfico 3.17 - Ensino (Cluster 2)	50
Gráfico 3.18 - Rendimento (Cluster 2)	50
Gráfico 3.19 - Faixa Etária (Cluster 3)	52
Gráfico 3.20 - Empregabilidade (Cluster 3)	53
Gráfico 3.21 - Ensino (Cluster 3)	53
Gráfico 3.22 - Rendimentos (Cluster 3)	53
Gráfico 3.23 - Faixa Etária (Cluster 4)	55
Gráfico 3.24 - Empregabilidade (Cluster 4)	56
Gráfico 3.25 - Ensino (Cluster 4)	56
Gráfico 3.26 - Rendimentos (Cluster 4)	56

## Lista de Tabelas

Tabela Anexo IV.1 - Dados provenientes de [11] no formato tabular (Excel) e com resolução espacial ao nível do município .....	155
Tabela Anexo IV.2 - Dados do tipo pontual que se encontram no sistema de referência WGS 1984, provenientes de [43] .....	156
Tabela Anexo V.1 - Dados utilizados na criação do Dashboard de Monitorização dos Cuidados de Saúde a nível de ARS e ACES .....	157

## Lista de Abreviaturas, Siglas e Acrónimos

ACES – Agrupamentos de Centro de Saúde  
ARS – Administrações Regionais de Saúde  
BGRI – Base Geográfica de Referenciação de Informação  
CAOP – Carta Administrativa Oficial de Portugal  
CSP – Cuidados de Saúde Primários  
DICOFRE - Código numérico do INE que identifica DI(strito), CO(ncelho) e FRE(guesia).  
ESRI – Environmental Systems Research Institute  
INE – Instituto Nacional de Estatística  
SNS – Sistema Nacional de Saúde  
SIG – Sistemas de Informação Geográfica

## Lista de Termos

*Feature* – objeto do mundo real representado num mapa, caracterizado por geometria (ponto, linha ou polígono) e atributos alfanuméricos.  
*Feature Class* – conjunto de objetos com o mesmo tipo de geometria, o mesmo conjunto de atributos e a mesma referência espacial que se organizam num tema de informação de um SIG.  
Serviço Web – conjunto de dados geográficos expostos pela web, através de um URL, para ser consumido em mapas e aplicações na internet.

# 1. Introdução

## 1.1 Enquadramento do Projeto

Os sistemas de informação georreferenciada são um instrumento prático para o apoio nos cuidados de saúde uma vez que através destes podemos compreender as desigualdades que existem num determinado território, na gestão da informação e mesmo em tomadas de decisão. Estes permitem capturar, armazenar, lidar e integrar geograficamente grandes volumes de informação provenientes de variadas fontes, programas e setores. Uma identificação do padrão espacial de uma doença faz com que as causas subjacentes possam ser identificadas e obter os dados e informações necessárias para poder aprovisionar e oferecer informações para combater as necessidades e eventual propagação [1][2].

Foi neste âmbito que foi desenvolvido um projeto na empresa “ESRI Portugal”, distribuidora exclusiva da empresa norte-americana ESRI, sendo esta líder na resolução de desafios na área dos Sistemas de Informação Geográfica. Esta é uma sucursal de uma empresa internacional de origem norte americana que foi fundada em 1987, atuando como agente especializado no desenvolvimento e fornecimento de sistemas de informação baseados na tecnologia de Sistemas de Informação Geográfica (SIG). O projeto em causa diz respeito à caracterização da população em Portugal, utilizando técnicas de segmentação como ferramenta para auxiliar a análise espacial de atributos relacionados com a área da saúde. respeito à caracterização da população em Portugal, utilizando técnicas de segmentação como ferramenta para auxiliar a análise espacial de atributos relacionados com a área da saúde. Esta ferramenta foi aplicada para entender a distribuição de uma doença crónica, a Diabetes em Portugal Continental. Entende-se segmentação de população como conceito para ajudar a entender as necessidades desta para que, por exemplo, serviços sejam melhor planeados e entregues [3]. A análise de autocorrelação espacial foi também utilizada neste trabalho, através da aplicação do índice Anselin Local Moran [4], que identifica regiões cuja semelhança dos atributos nos locais onde se encontram realçam padrões espaciais e *outliers* de interesse e que permitiu de certa forma compreender os processos gerados pela segmentação.

Para informação adicional, e de forma a mostrar também o panorama dos cuidados de saúde em Portugal foram criados vários produtos de informação, nomeadamente *Operations Dashboard* e um *Visualizador*. Um *Operations Dashboard* é uma aplicação web configurável que permite visualizar dados espaciais e respetivas análise dos mesmos permitindo uma visualização operacional em tempo real de pessoas, serviços ativos e eventos. Através de um painel dinâmico é possível visualizar atividades e principais indicadores de desempenho consoante o objetivo deste mesmo produto. O *Visualizador* é também uma aplicação web configurável só que, no entanto, é composto por várias aplicações denominadas de *widgets*, que possibilita um número de possibilidade para analisar e visualizar dados espaciais.

A diabetes é a oitava causa de morte mais comum no mundo sendo uma das maiores problemáticas do século XXI [5]. É cada vez mais diagnosticada em populações mais jovens, sendo que a prevalência mundial aumenta de ano para ano, uma vez que estas têm uma alimentação cada vez mais calórica acompanhada de poucas práticas de exercício físico [6].

Em 2017, a nível mundial, cerca de aproximadamente 427 milhões de pessoas com faixa etária compreendida entre os 20 e 79 anos são diabéticas, sendo que é esperado um aumento de cerca de 48% deste número até 2045. Cerca de 79% dos adultos diabéticos vivem em países com baixo-médio rendimentos, ou seja, verifica-se uma predominância desta pandemia em situações socioeconómicas desfavorecidas [7].

Cerca de 12% dos gastos na área da saúde são devido aos diabetes, sendo que um em cada quatro dólares americanos que são gastos a nível mundial relativamente a cuidados de saúde para diabetes são gastos apenas na Europa [7].

Em Portugal, em 2015, a prevalência estimada de diabetes com faixa etária compreendida entre os 20 e 79 anos (7,7 milhões de indivíduos) foi cerca de 13,3%, e foi estimado que cerca de 44% da população portuguesa não estava diagnosticada [8], [9]. Aliado a este diagnóstico, cerca de 2 milhões de pessoas têm hiperglicemia intermédia, estando em risco de virem a desenvolver diabetes [10].

No entanto, através do gráfico 1.1 verifica-se que a taxa de prevalência aumenta com a idade e que o sexo masculino apresenta uma maior prevalência que o sexo feminino.

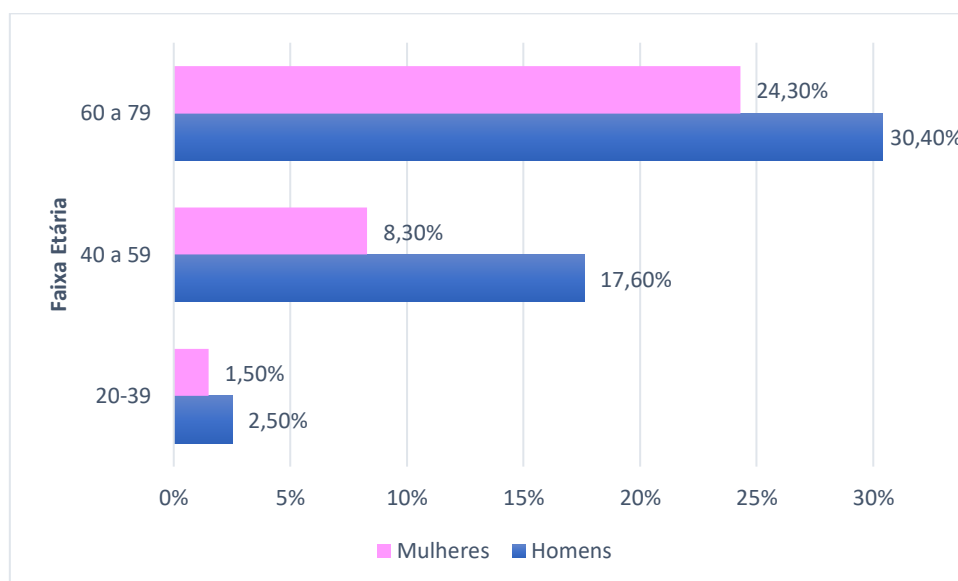


Gráfico 1.1 - Taxa de Prevalência de diabetes por Faixa Etária e Género. (Extraído de [9]).

Relativamente à incidência de diabetes (gráfico 1.2), verificamos uma tendência de crescimento, mesmo após os anos de 2010 e 2011, onde se verificaram os valores máximos de novos casos.

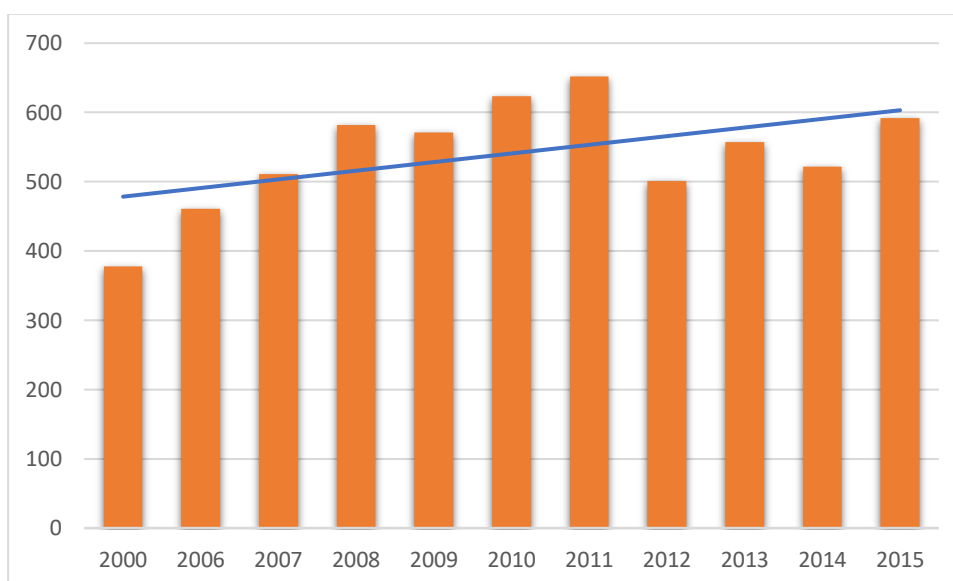


Gráfico 1.2 - Número de novos casos de diabetes por 10 000 indivíduos por ano. (Extraído de [9])



O ano em que existiu um maior número registado de óbitos devido a diabetes foi em 2012, como se pode verificar no gráfico 1.3, e este número tem vindo a diminuir, o que é confirmado através da expressão da linha de tendência, pois segundo o gráfico 1.3 é esperado que ao longo do tempo o número de óbitos, devido a diabetes, diminua.

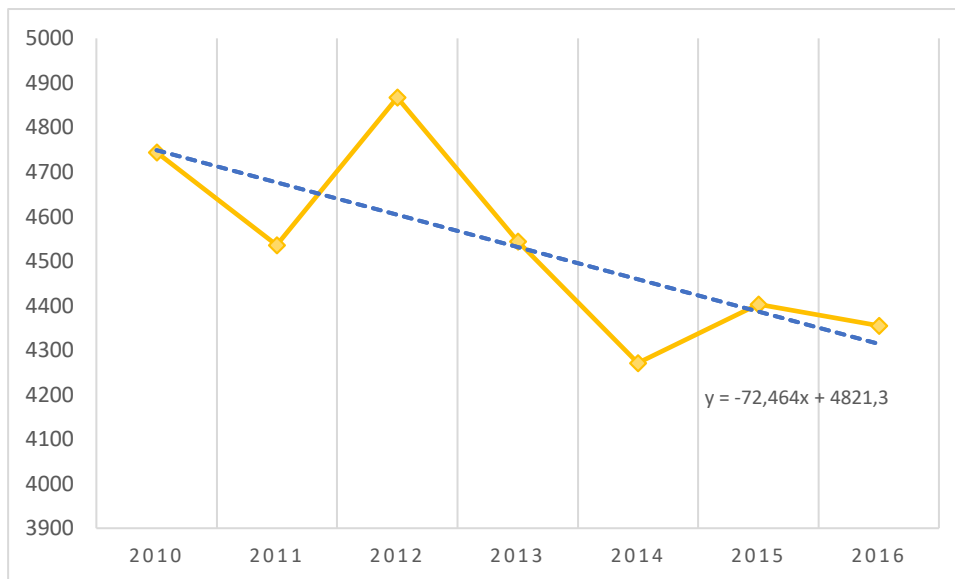


Gráfico 1.3 - Número de Óbitos devido a diabetes em Portugal por ano. (Extraído de [11])

Comparando o gráfico 1.3 com o gráfico 1.4, verificamos que as tendências são inversas, ou seja, apesar do número de óbitos em Portugal aumentar, através do gráfico 1.3 podemos verificar que o número de óbitos devido a diabetes está a diminuir.

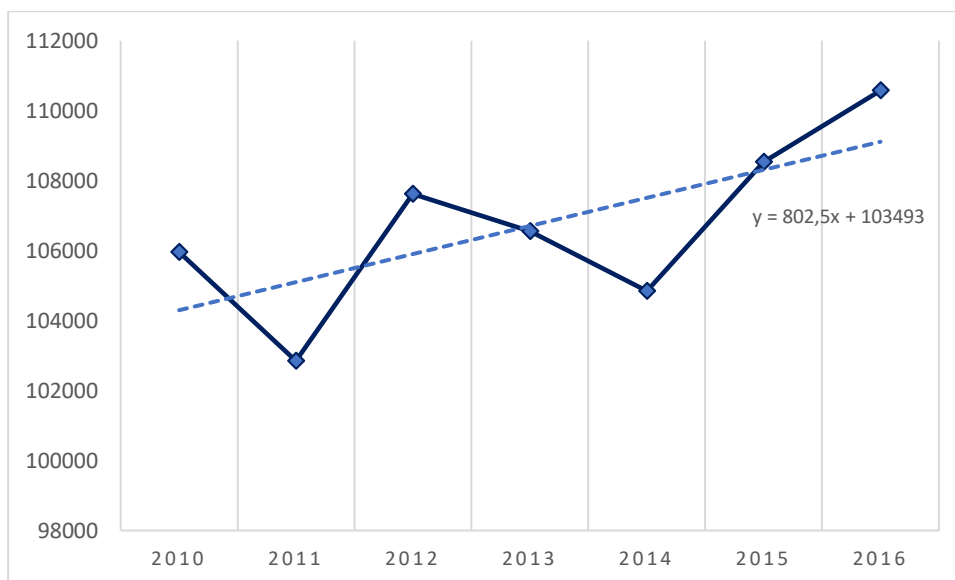


Gráfico 1.4 - Número total de óbitos em Portugal por ano. (Extraído de [11])

Para 2020, o Programa Nacional para a Prevenção e Controlo da diabetes traçou três metas: Evitar o desenvolvimento de diabetes em cerca de 30 000 utentes de risco identificado, aumentar em 30 000 o número de novos diagnósticos de diabetes através de um teste precoce e diminuir em 5% a mortalidade por diabetes antes dos 70 anos [10].

Verifica-se uma estagnação na evolução de indicadores e uma diminuição da respetiva cobertura assistencial à população diabética registada nos cuidados de saúde primários, o que indica que é necessário existir um acompanhamento e monitorização nesta área [9].

Os cuidados de saúde primários são o pilar central do sistema de saúde, uma vez que tem responsabilidades na promoção da saúde, prevenção da doença e de prestação de cuidados de saúde. As unidades prestadoras de cuidados de saúde primários estão integradas em Agrupamentos de Centros de Saúde (ACES) e em Unidades Locais de Saúde (ULS), sendo que o foco deste projeto está nos ACES [10].

Os ACES são serviços públicos de saúde que têm autonomia administrativa com a missão de garantir a prestação de cuidados de saúde primários à população de uma determinada área geográfica. Em Portugal, existem cerca de 55 ACES [10]. Vários ACES encontram-se distribuídos em áreas regionais de saúde (ARS), sendo que em Portugal existem cinco: Algarve, Alentejo, Lisboa e Vale do Tejo, Centro e Norte.

Como afirmado anteriormente, os SIG são uma boa ferramenta para efetuar este tipo de monitorização, uma vez que ao verificarmos a distribuição espacial desta patologia, podemos analisar qual a influência de variados atributos, desde demográficos, socioeconómicos, entre outros, e a partir dessa análise, implementar políticas de forma a mitigar estas influências e reduzir o número de incidências da patologia. Esta é uma tecnologia versátil e bastante útil para a área da Saúde, uma vez que para além de fornecer técnicas e ferramentas para analisar espacialmente uma doença, também permite caracterizar e verificar tendências de utilização dos serviços, tendo em conta, por exemplo, características da população que acede a estes, entre uma variedade de utilizações que serão discutidas mais à frente neste relatório.

## 1.2 Objetivos

De forma a promover políticas de saúde mais eficazes e para uma melhor gestão dos recursos, é necessário compreender de que forma a saúde humana pode ser afetada por diversos fatores, nomeadamente socioeconómicos, também com o intuito de mitigar possíveis ameaças para a saúde dos cidadãos.

Pretende-se com este estudo, e tirando máximo partido da tecnologia ESRI, gerar uma segmentação da população em Portugal e aplicar ferramentas de análise espacial para explorar diversas variáveis e compreender qual o impacto destas sobre uma doença crónica, como a diabetes. Será também feita uma análise da *performance* dos CSP em Portugal.

Este projeto tem ainda como objetivo a criação de quatro produtos de informação

- *Operational Dashboard* com resultados da segmentação e dados da diabetes – Este produto tem como objetivo mostrar como se distribuem os grupos gerados através da segmentação e qual o desempenho dos atributos utilizados, sendo que é também incluído dados relativo à diabetes de forma a demonstrar também como estes se encontram numericamente distribuídos pelo espaço.

- *Operational Dashboard* de monitorização dos cuidados de saúde a nível de ARS e ACES – Qual o estado dos CSP de Portugal? Este produto tem como objetivo mostrar todos os indicadores disponíveis que demonstrem a característica dos ACES e ARS a nível individual ou nacional, quer seja em termos de profissionais de saúde ou de características dos serviços que disponibilizam;

- *Operational Dashboard* de monitorização/gestão proativa dos cuidados de saúde a nível de Municípios – Este produto permite caracterizar os serviços e apoios ao nível de infraestruturas de Saúde em Portugal, para os municípios, seja a quantidade de pessoal ao serviço, o número de instituições de saúde, e é feita uma comparação de indicadores de Centros de Saúde comparativamente aos de Hospital;

- Visualizador com áreas de influência de Hospitais em Lisboa – Através deste produto será possível verificar quais as áreas de influência dos Hospitais em Lisboa, e quais as instituições de Saúde, nomeadamente Hospitais, Centros de Saúde e Farmácias mais perto de qualquer localização desejada;

## 1.3 Estrutura do Relatório

Este relatório encontra-se dividido em 5 grandes capítulos. O primeiro capítulo faz uma introdução e contextualização do projeto, sendo que também apresenta os objetivos deste mesmo.

O capítulo 2 aborda alguns conceitos teóricos, ferramentas e metodologias utilizadas em SIG, e também como podem essas metodologias ser aplicadas na área da saúde. São debatidos temas como as várias aplicações dos SIG na área da saúde, que aplicabilidade tem o conceito de segmentação nesta área e é descrito de forma resumida a diabetes e a utilização dos SIG para estudo desta doença.

Em seguida, o terceiro capítulo é composto por toda a componente de aquisição, pré-processamento de dados e metodologia aplicada para a obtenção dos resultados da segmentação da população ao nível de freguesias. Apresenta também metodologia e os resultados da autocorrelação espacial local, sendo que este termina com a análises subsequentes da segmentação e análise espacial.

O quarto capítulo é referente aos produtos de informação criados, onde são descritos todos os dados utilizados na criação deste, sendo que também é mostrado a metodologia e o resultado final, onde consta o aspeto visual dos produtos de informação criados como também qual o endereço em que os podem consultar.

Para finalizar, o capítulo 5 apresenta as conclusões deste estudo e algumas recomendações relativas à possibilidade de utilizar a metodologia deste projeto em estudos semelhantes no futuro. Este documento apresenta ainda um conjunto de anexos onde se pode encontrar todos os mapas gerados através do software utilizado e que são utilizados também no corpo do texto deste relatório em tamanho mais reduzido.

## 2. Fundamentos Teóricos e Estado de Arte

### 2.1 SIG na área da Saúde

Os cuidados de saúde são um grande e complexo sector de economia no panorama mundial. Qualquer país tenta fornecer os melhores cuidados de saúde e com o menor custo possível aos seus cidadãos. Ainda existem sistemas informáticos para a saúde que são manuais ou não integrados, ou seja, organizações tem vários sistemas de informações que não se encontram interligados, o que leva a um grande custo nas transações, erros médicos e o dobro de esforços, como por exemplo, efetuar o mesmo teste médico mais que uma vez [12].

A importância dos serviços informáticos na área da saúde tem vindo a aumentar ao longo dos anos, uma vez que estes beneficiam o planeamento, gestão, pesquisa e a criação de políticas nesta área. Cada vez mais os sistemas tornam-se integrados e automatizados, beneficiando as aplicações na área dos SIG, pois irá existir um maior volume de dados, seja informações sobre pacientes, programas, incidência de doenças, estando agregado a estes dados das respetivas localizações, enriquecendo assim estas aplicações [12].

As capacidades analíticas dos SIG fazem com que na área da saúde pública exista a oportunidade de examinar as relações entre as características da saúde da população e as características ambientais tanto humanas como físicas, como se pode ver na figura 2.1 [13].



Figura 2.1 - SIG e algumas variáveis

Na área da saúde pública e promoção de cuidados de saúde, os SIG são utilizados principalmente a nível de monitorização de doenças, análise de risco, planeamento e acesso a cuidados de saúde e o perfil de saúde da comunidade [2][14].

#### 2.1.1 Monitorização de Doenças

Na área de monitorização de doenças, os dados têm em conta a prevalência, incidência e a propagação da doença, sendo que se recorre a esta ferramenta para entender a distribuição e propagação geográfica da doença. Existem duas componentes nesta área, o mapeamento da doença e a modelação desta [14].

O mapeamento serve para visualizar e compreender o comportamento da doença no passado e presente. O seu objetivo principal é a prevenção de doenças, sendo que acaba por oferecer suporte a nível do controlo de riscos. O foco principal é a deteção de *clusters* espaciais da doença. Um *cluster* espacial é o conjunto de duas, ou mais, áreas que se encontram circunjacentes e que partilham similaridades em termos de atributos. Esta deteção facilita a avaliação dos resultados dos cuidados de

saúde, verificação dos riscos na população, construção de hipóteses e o planeamento de medidas de prevenção e na apresentação de indicadores de saúde.

A modelação é uma extensão do mapeamento sendo que com esta podemos compreender a propagação da doença, prevendo a sua expansão espacial futura, identificando fatores de risco para a transmissão e verificar quais são as áreas de maior risco para prevenção e intervenção. O fato de ser adicionada informação estatística espacial ou outras técnicas analíticas ao mapeamento faz com que seja possível gerar modelos que prevêem padrões de propagação da doença. A grande vantagem reside no fato de ao representar a distribuição espacial da doença é possível descobrir e explorar relações entre atributos de saúde com lugares que poderiam não ser perceptíveis com técnicas tradicionais. Esta ferramenta é mais indicada para doenças vetoriais ou sexualmente transmissíveis [14].

Um problema que ocorre ao tentar inferir as causas de um surto de doença e formular uma resposta ao nível dos cuidados de saúde a partir de *clusters* espaciais detetados é o fato da componente temporal não ser incluída adequadamente, pois a localização da causa de exposição e sua respetiva manifestação raramente coincidem devido a desfazamento de tempo-espço, como a mobilidade da população e taxa de exposição [2].

Um estudo feito pela Universidade Federal da Paraíba [15] teve com objetivo analisar a distribuição espacial e os focos de Hanseníase, mais conhecido como Lepra, no município de Bayeux, no estado do Paraíba, entre 2001 e 2011. Os dados utilizados foram os casos diagnosticados de Lepra fornecidos pelo Centro de Controlo de Hanseníase, sendo que estes tinham como atributos: o endereço completo, o posto de saúde de origem, sexo, idade, número de residentes no mesmo domicílio, estado civil, número de lesões, grau de incapacidade, forma clínica e se era gestante. Através da figura 2.2, é possível visualizar a distribuição das Unidades de Tratamento de Hanseníase (UTH) e a densidade de casos desta doença para o período dos dados.

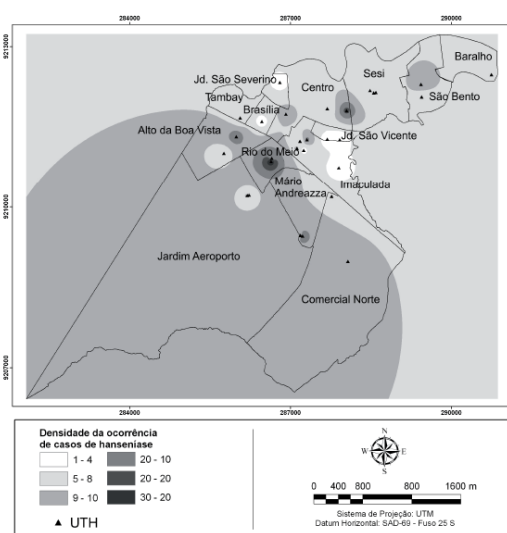


Figura 2.2 - Distribuição das Unidades de Tratamento de Hanseníase (UTH) e a densidade de casos desta doença para o período dos dados. (Extraído de [15])

As UTH com maiores concentrações de caso registrados são as unidades localizadas nos bairros de rio do Meio, Mário Andreazza, Alto da Boa Vista e Centro. Este mapa representa as áreas de maior concentração da doença onde o risco de adoecimento é provavelmente maior. Os fatores relativos à distribuição espacial desta doença podem-se agrupar em naturais, como por exemplo, o clima, o relevo, tipos de vegetação e determinados ecossistemas, e sociais, como condições desfavoráveis de vida,

desnutrição e movimentos migratórios. Na figura 2.3 observa-se que 26 casos da doença se situam em áreas de ocupação subnormal. Um exemplo deste tipo de ocupação são as favelas. Os autores afirmam que o espaço socialmente organizado, integrado e profundamente desigual possibilita e determina a ocorrência de endemias e a sua propagação. Tendo este pensamento, os autores notaram uma pequena relação entre a concentração de casos da doença em aglomerados subnormais neste município.

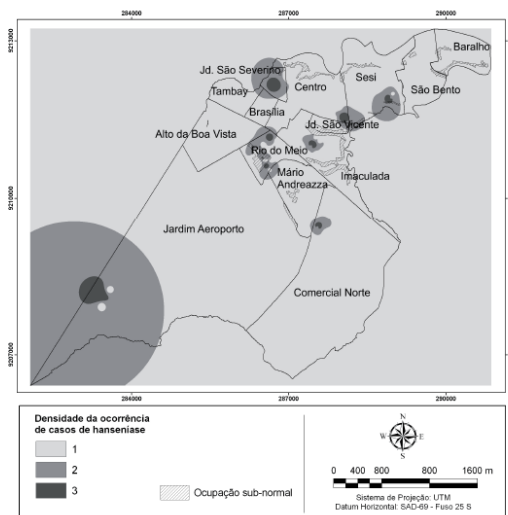


Figura 2.3 - Localização das ocorrências de hanseníase e áreas de ocupação subnormais em Bayeux. (Extraído de [15])

Os autores consideraram as técnicas de geoprocessamento valiosas e eficazes para o entendimento epidemiológico e na ordenação de ações para diminuir a expansão da doença. A visualização espacial da endemia providenciou a informação para aferir a localização real dos casos, mostrando que a doença se localiza mais em regiões onde reside a população com baixo padrão socioeconômico. A partir destas técnicas também conseguiram verificar focos da endemia que deverão ser trabalhados a partir de políticas de controle.

## 2.1.2 Análise de Risco

A área da análise de risco encontra-se centrada nos fatores ambientais. Ou seja, pretende analisar o impacto desses mesmos fatores na saúde pública. Neste caso é feita a avaliação, a comunicação, a gestão e a monitorização do risco desse mesmo fator. Por exemplo, qual é o risco de propagação de uma doença se a população se encontrar nas proximidades de uma zona industrial? O principal objetivo destas análises é identificar quais as necessidades e aumentar a eficiência do controle para prevenir epidemias ou surtos de uma doença. Normalmente existe apenas uma componente de análise e avaliação e não de monitorização ou gestão desses mesmos fatores. [14]

Para obter o melhor contexto possível na resolução de um problema é necessário considerar variáveis que representem a população e o ambiente adequadamente. Inicialmente é necessário verificar qual é componente espacial do fenómeno que está a ser estudado e que relações espaciais existem, por exemplo, a população encontrar-se centrada em determinado lugar. [14]

Um estudo feito pelo Departamento das Ciências Geológicas, Geográficas e ambientais da Universidade de Nova Iorque [16] teve como objetivo fazer uma análise de risco de poluição no ar e ver que relação espacial teria com a asma no condado de Bronx. Os dados necessários para conduzir esta análise foram os registos de hospitalização por asma, a localização e a emissão de informações sobre instalações responsáveis pela poluição e informações demográficas e socioeconómicas. Recorrendo à

tecnologia SIG, verificaram que minorias e populações pobres tinham maior probabilidade de se encontrar concentradas em, ou perto de, zonas industriais, como se pode ver na figura 2.4.

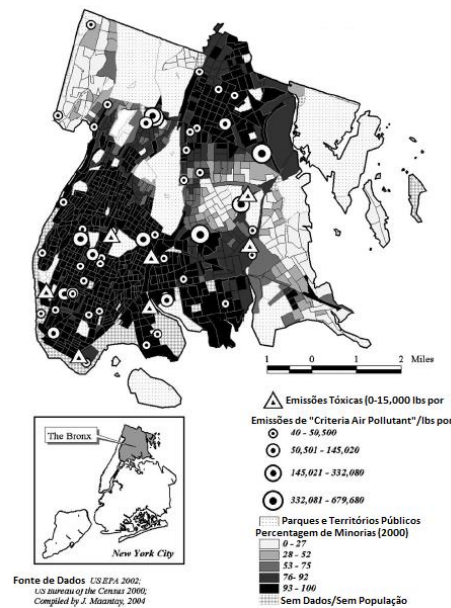


Figura 2.4 - Principais fontes estacionárias de poluição e população minoritária em Bronx. (Adaptado de [16])

Em Nova Iorque cerca de 22% da população residente vive perto de grandes zonas industriais, considerando o impacto ambiental e a nível de saúde destas estruturas algo a ter em consideração. Para os dados relativo à hospitalização devido a Asma, foram tidos em conta médias de 5 anos, divididas por total de população, adultos e crianças, figura 2.5.



Figura 2.5 - Hospitalização devido a Asma. (Adaptado de [16])

Duas formas comuns de poder determinar o potencial de exposição é o método de coincidência espacial e análise de proximidade. O método de coincidência espacial tem em conta a caracterização da população dentro de uma unidade geográfica (seja censitária, código postal, freguesia, entre outros) e verificar se existe uma instalação poluidora nessa unidade. Populações que se encontrem em unidades que contenham esse tipo de instalações são consideradas impactadas por elas e potencialmente expostas a efeitos nefastos. As que não se encontram nessas unidades não são consideradas impactadas pelas instalações. É considerada uma análise simplista e uma forma imprecisa de avaliar este tipo de impacto em populações. Por exemplo, há quem possa viver na mesma unidade geográfica que uma destas instalações, no entanto viver longe desta, mas segundo este método será considerado impactado pela instalação, como há quem possa viver numa freguesia ao lado de uma instalação, mas sendo assim já não é considerado impactado.

A análise de proximidade examina a população que se encontra a uma distância específica da instalação poluidora. Esta distância é calculada tendo em conta o tipo de instalação e o tipo de emissões provenientes desta. A população que se encontra a uma determinada distância, esta criada através de um buffer, é considerada impactada e a que se encontrar depois já não o é. Um exemplo desses buffers criados tendo em conta o tipo de fonte de poluição encontra-se na figura 2.6.

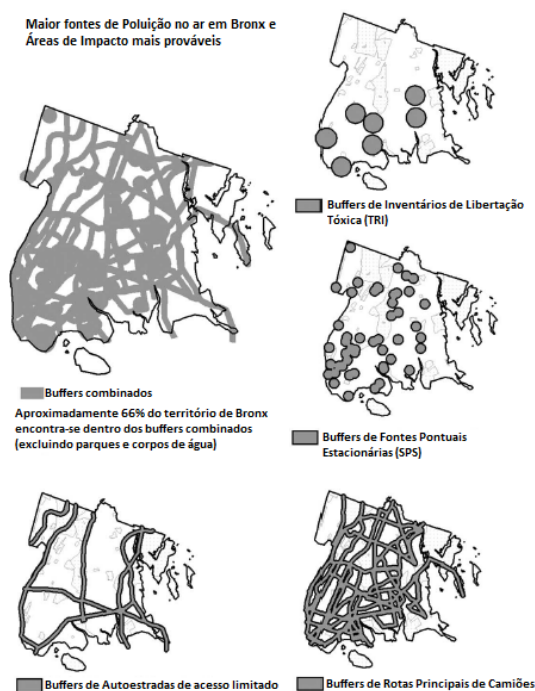


Figura 2.6 - Buffers de Proximidade de Fontes de Poluição. (Adaptado de [16])

Contudo, esta análise de proximidade também tem falhas, uma vez que assume que toda a população dentro do buffer é impactada, o que não corresponde à realidade uma vez que a poluição não se dispersa uniformemente em todas as direções no ar. As distâncias utilizadas para a construção dos buffers são baseadas em padrões de qualidade ambiental já estabelecidos, evidências empíricas sobre o transporte de poluentes, e resultados de modelos gerais. Graças a esta técnica conseguiram entender que cerca de 66% do território do condado se encontrava dentro dos buffers e que maioria da população se encontraria exposta aos poluentes e que era 30% mais provável de uma pessoa que se encontra dentro da área dos buffers ser hospitalizada devido a asma do que uma fora.



### 2.1.3 Planeamento e Acesso a Cuidados de Saúde

Quanto ao planeamento e acesso a cuidados de saúde, neste tipo de aplicações é efetuado um estudo que descreve a capacidade da população de usufruir os cuidados de saúde quando precisam, ou seja, estuda as relações entre as variáveis que descrevem as necessidades da população e aquelas que os serviços de saúde oferecem. Neste tipo de estudos o foco está na acessibilidade espacial e temporal dos indivíduos aos serviços de saúde. [14]

Por exemplo, construir os perfis dos indivíduos que utilizem um determinado serviço de saúde permite a esse mesmo serviço melhorar e entender a necessidade de quem recorre a ele e maximizar a eficiência da utilização dos recursos do serviço. Em termos de acessibilidade também é importante ter em conta a localização dos serviços de saúde, seja a distância ou a capacidade de viajar entre eles, criando assim uma rede de serviços de saúde. Ao efetuar uma análise desta rede pode-se verificar qual o potencial de uma área tendo em conta a população que consegue alcançar e quais as suas características.

### 2.1.4 Construção do Perfil de Saúde da Comunidade

O perfil de saúde da comunidade acaba por ser a junção das outras três categorias descritas anteriormente, monitorização de doenças, análise de risco, planeamento e acesso a cuidados de saúde. Estes perfis podem incluir dados que influenciem a saúde de forma direta ou indireta, tais como, características sociodemográficas, mortalidade, comportamentos de saúde (alcoolismo, tabagismo, práticas sexuais), políticas (legislação de pesticidas, proibição de consumos). Estas variáveis estão ligadas à localização espacial de infraestruturas da comunidade, tais como, igrejas, escolas, hospitais, restaurantes, supermercados, estradas, entre outros. Esta ligação permite verificar quais são as relações gerais entre o cenário da comunidade e a saúde.

Estes perfis podem ser usados para variadas soluções, por exemplo, para verificar mudanças na saúde de uma comunidade ao longo do tempo, fazer sugestões para uma análise futura ou para desenvolver hipóteses e obter mais informações da comunidade. Uma vez que doenças e saúde vão se alterando consoante a localização ou o tempo, é necessário que os agentes de cuidados de saúde possam ter uma ferramenta que seja eficiente e consiga visualizar as necessidades das comunidades. É neste contexto que a segmentação da população se torna útil, uma vez que grupos serão formados a partir de características comuns, esta temática será discutida no subcapítulo 2.4.

## 2.2 Diabetes

A diabetes é uma doença crónica, caracterizada por um grupo complexo de distúrbios metabólicos, no qual a quantidade de açúcar no sangue, denominada também por glicose, é muito elevada devido ao pâncreas não produzir qualquer ou pouca insulina, sendo que a insulina produzida poderá não funcionar adequadamente [6], [17].

A insulina é uma hormona que é produzida pelas células  $\beta$  do pâncreas e é a que possibilita a entrada da glicose nas células, caso esta se encontre em défice, a glicose acumula-se no sangue em vez de entrar nas células, fazendo com que estas não funcionem de forma correta.

Num corpo saudável, quando a quantidade de glicose aumenta no sangue, o pâncreas liberta quantidade suficiente de insulina para a circulação sanguínea de modo a estabilizar os níveis de glicémia, concentração de glicose no sangue, e permitir a entrada da glicose nas células para estas a utilizarem como energia, como se pode verificar na figura 2.7 [18].

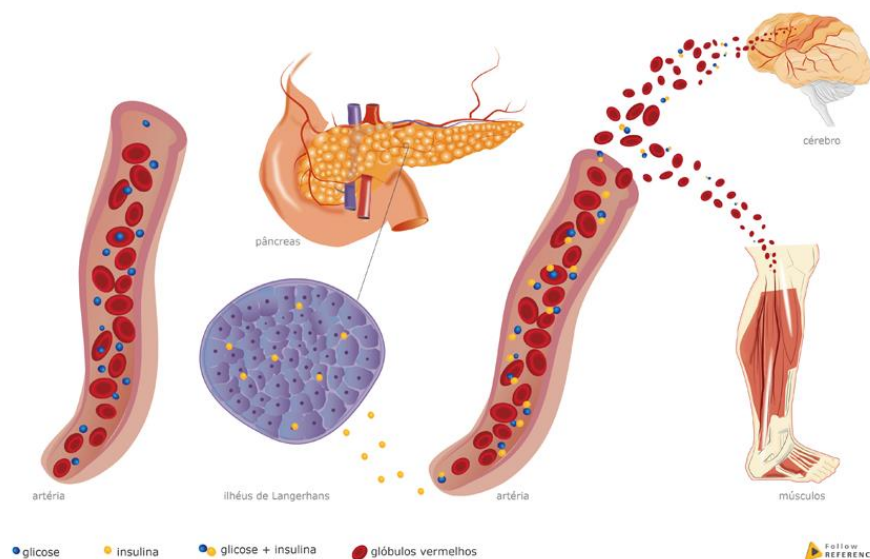


Figura 2.7 - Insulina a atuar num organismo saudável. (Extraído de [18].)

Apesar de existirem vários tipos desta doença, existem dois principais: diabetes tipo 1 e tipo 2. A diabetes tipo 1 é aquela que é mais frequente em crianças e adolescentes [6].

Neste caso o sistema imunitário ataca e destrói as células  $\beta$  do pâncreas. Este sistema por vezes funciona de forma anormal, atacando e destruindo partes do organismo do ser humano, daí que há quem denomine a diabetes do tipo 1 como uma doença autoimune, e por norma é uma doença que resulta de heranças hereditárias. Neste tipo de doenças, o portador é considerado insulino dependente pois necessitam de medicação com insulina durante toda a sua vida [6], [18].

Neste caso, uma vez que é o próprio sistema imunitário que destrói as células produtoras de insulina, o organismo não irá ter insulina para poder reduzir os níveis de glicémia, fazendo com que a glicémia permaneça na corrente sanguínea e não entre nas células, figura 2.8.

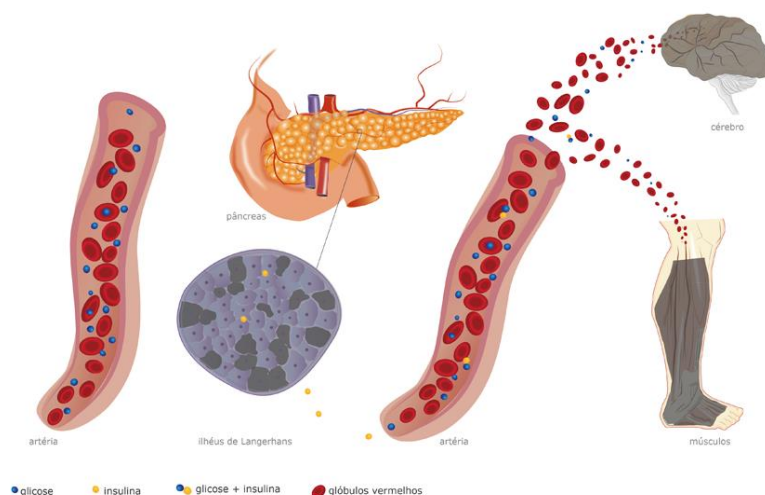


Figura 2.8 - Insulina a atuar num organismo portador de diabetes tipo 1. (Extraído de [18])

A diabetes tipo 2, figura 2.9, é o tipo mais comum de diabetes. Normalmente encontra-se em pessoas adultas, pois está associada a maus hábitos de vida, seja relacionado sedentarismo ou

alimentação, no entanto existe também uma componente hereditária. Em Portugal, cerca de 1 em cada 10 portugueses tem este tipo de diabetes.

Neste caso, a insulina encontra-se em défice e existe uma resistência a esta, sendo a resistência a perda progressiva da eficácia da insulina. Isto significa que numa fase inicial, se o organismo é resistente à insulina, este poderá apresentar valores elevados de insulina e de glicose normais. Mas com o passar do tempo, o organismo vai deixar de conseguir compensar este desequilíbrio, existindo uma diminuição anormal e progressiva da produção de insulina, fazendo os níveis de glicose na corrente sanguínea aumentar [18].

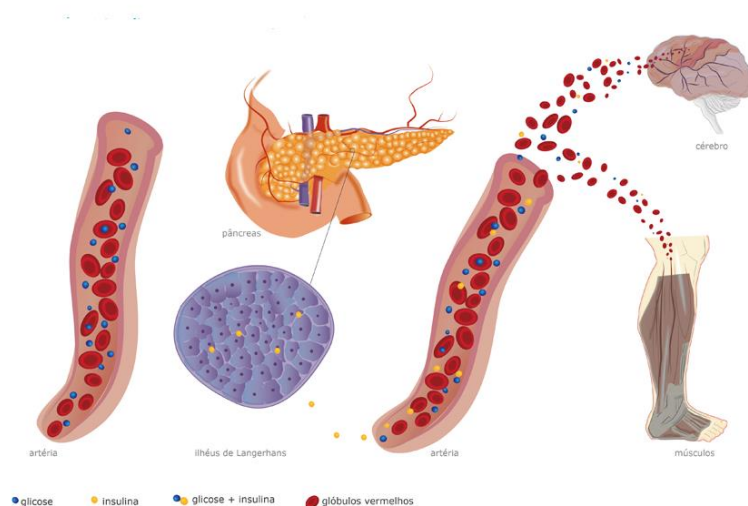


Figura 2.9 – Insulina a atuar num organismo portador de diabetes tipo 2. (Extraído de [18])

Para existir uma boa prevenção e controlo desta doença é necessário que o indivíduo portador conheça bem a doença, que tenha bons hábitos de alimentação, com uma dieta equilibrada e variada, controle os níveis de glicose no sangue através de exames periódicos e que efetue devidamente o tratamento prescrito pelo seu médico.

É possível controlar a diabetes através de dois exames, um feito ao pé e outro feito para controlar a hemoglobina glicada, denominada por HbA1c. O exame ao pé é uma inspeção periódica anual em que são identificados fatores de risco condicionantes de lesões dos pés, seja em termos de úlceras, ou a avaliação do estado das unhas e da pele, e consoante o risco avaliado ao utente é delineado um tratamento. O exame aos níveis de HbA1c deverá ser feito semestralmente às pessoas com diabetes, sendo que este cada vez mais é considerado um instrumento de monitorização em programas de avaliação da qualidade dos cuidados prestados pelos profissionais de saúde nesta área [19].

## 2.3 Diabetes e SIG

Hoje em dia já existe bibliografia variada a nível mundial que relacione a diabetes com os SIG, no entanto em Portugal é ainda uma área pouco explorada.

As ferramentas SIG e as funcionalidades analíticas disponíveis nestes sistemas podem apoiar profissionais de cuidados de saúde primário quando se pretende analisar com estas metodologias a grandes bases de dados de registos médicos disponíveis em formato digital. Há bibliografia que comprova que a condição socioeconómica era uma barreira para fazer um bom controlo de glicose, mas não para fazer um controlo de lípidos em indivíduos portadores de diabetes tipo 2, sendo que bairros

onde existisse um menor valor de rendimentos iria existir maiores níveis de glicose, o que significa um controlo menor desta doença [20].

Num estudo que tiveram em conta dados de diabetes em crianças e adolescentes na Suécia [21], ao efetuarem o mapeamento destes mesmos dados e recorrendo ao método de regressão Kriging para estimar a distribuição espacial da incidência de diabetes, sendo que concluíram que existia uma variação geográfica significativa nas taxas de incidência entre municípios, a maior resolução espacial, figura 2.10.

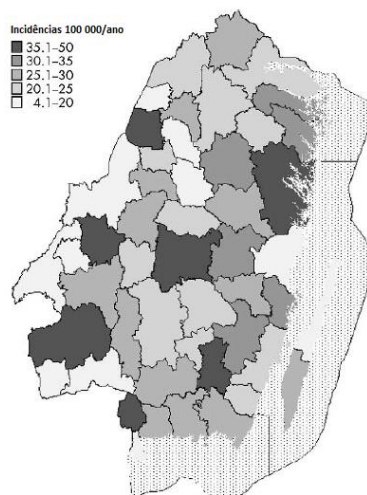


Figura 2.10 - Número de Incidências de diabetes na Suécia. (Adaptado de [21])

Mas ao nível de regiões, figura 2.11, a variação geográfica dessa mesma taxa já não era significativa. Caso retirassem as 6 maiores cidades da área de estudo, sendo estas maioritariamente zona urbana, a variação entre municípios tornava-se mais fraca. Nos municípios com um maior risco, as maiores contribuições vinham de crianças com uma faixa etária de 6 a 10 anos, sendo que indivíduos do sexo masculino e feminino tinham a mesma variação geográfica. Acabam por concluir que a explicação mais provável para esta variação no risco de crianças e adolescentes desenvolverem diabetes tipo 1 seria que fatores ambientais locais teriam influência no processo de desenvolver esta doença.

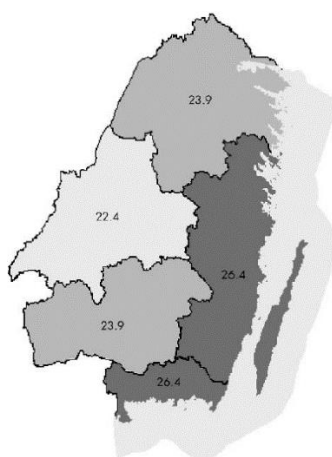


Figura 2.11 - Taxa de Incidência de diabetes nos 5 condados da zona sudeste da Suécia. (Extraído de [21])

Na Índia também foi realizado um estudo que recorreu aos SIG para tentar entender a influência do ambiente que rodeava os indivíduos portadores de diabetes e fatores de risco, fazendo cálculos de distância entre a residência do indivíduo e pontos de interesse, como entidades de saúde,

estabelecimentos que vendem comida e álcool, ou seja, verificar a acessibilidade. Também utilizaram ferramentas de agregação espacial, para entender qual o número de estabelecimentos que vendessem comida ou o número de parques, ferramentas de *Clustering*, para entender se os *clusters* da doença têm tamanho geográfico suficiente para não terem ocorrido por acaso, entre outros [22].

Dados demográficos como a zona de residência, educação, rendimentos ou etnia são constantemente utilizados neste tipo de estudos. Em Michigan (EUA) foi efetuado um estudo que tinha como objetivo verificar quais os municípios do estado com grande prevalência de diabetes e pouco, ou nenhum, acesso a recursos médicos e, ou, de comunidade. Mapearam este tipo de dados recorrendo à tecnologia SIG, verificando que ao longo dos anos a prevalência ia aumentando, no entanto, o aumento na zona Norte do estado era maior que na sua contraparte. Ao analisarem os resultados concluíram que municípios com alta prevalência de diabetes e poucos recursos eram facilmente identificáveis. Por exemplo, através da figura 2.12, graças à tecnologia SIG, os autores conseguiram mapear zonas de intervenção, sendo que a população alvo seria indivíduos diabéticos com baixo rendimento [23].



Figura 2.12 - Identificação de populações alvo para intervenção na área da diabetes em Michigan. (Adaptado [23])

Em Portugal, já foi efetuado um estudo na área metropolitana do Porto, em que o seu objetivo principal foi a criação de um índice de privação para o risco de diabetes, figura 2.13. Como variáveis para a análise multicritério foi considerado a distribuição geográfica da prevalência de diabetes e variáveis socioeconómicas e ambientais disponíveis à freguesia, como por exemplo, dados dos censos de 2011 e características biogeofísicas do território através da carta oficial administrativa de Portugal (CAOP) e da carta de ocupação do solo (COS). O autor chegou à conclusão de que as freguesias cujos residentes tivessem um maior peso relativo de idosos, mulheres, menor grau de literacia, maior número de desempregados e os que habitam as áreas mais degradadas ambiental e socialmente coincidam com as freguesias de residência onde existia um maior número de internamentos e maiores custos com o tratamento da diabetes [24].

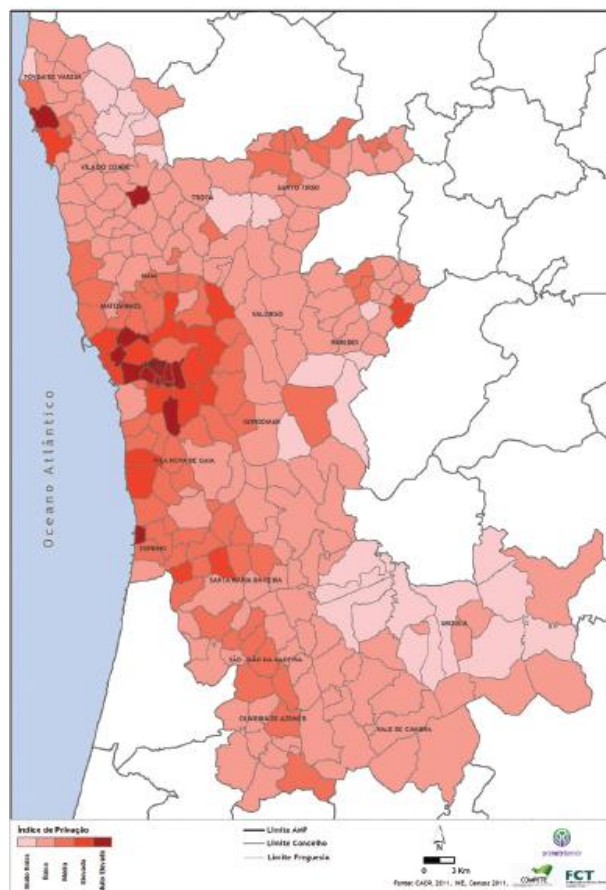


Figura 2.13 - Índice de Privação para o Risco de diabetes na área metropolitana do Porto (Extraído de [24])

## 2.4 Segmentação na área da saúde

Normalmente as pessoas encontram-se segmentadas por grupos, como por exemplo, adultos que trabalham e desempregados, religiosos e não religiosos, entre muitos outros exemplo, sendo que estes ainda podem ser subdivididos por classe social, etnia, nível de rendimento, uso de serviços públicos, entre outros [25].

Ao segmentar a população o principal objetivo é definir um número de grupos a partir dos quais todos os membros do mesmo grupo são o mais semelhante entre si e o mais dissemelhantes possível entre membros de outros grupos.

A forma de esta segmentação ser feita é através de um algoritmo de *Clustering* [25]. O *Clustering* é um processo de agrupamento de um conjunto de objetos físicos ou abstratos em classes ou objetos similares. É um método de aprendizagem não supervisionado uma vez que não sabemos a que grupos cada observação pertence. Este processo é usado em várias aplicações, tais como, análise de imagem, pesquisa de mercado ou segmentação de clientes. O resultado depende dos dados utilizados e de qual o algoritmo escolhido. Existem variados algoritmos de *Clustering*, tais como, hierárquico, partição ou baseados na densidade [26].

Neste projeto utilizou-se um algoritmo de partição denominado por K-means [27]. Este é um algoritmo que agrupa os dados ao tentar separar estes em grupos, minimizando o erro quadrático destes [26]. Este algoritmo necessita que o número inicial de *clusters*,  $n_c$ , seja fornecido e a métrica utilizada é a distância entre pontos, nomeadamente a distância euclidiana.

Por vezes, quem recorre ao algoritmo sabe o número mais apropriado de *clusters*  $n_c$  para o problema, no entanto, em muitos casos, o utilizador não tem um critério para escolher um número específico de *clusters*, apenas quer aquele que melhor distingue as similaridades e diferenças nos dados [28]. Não existe um método que determine um valor exato de  $n_c$ , mas sim uma estimativa precisa que pode ser obtida através de determinadas técnicas [29]. Neste projeto, foi utilizado o *Calinski-Harabasz pseudo F-statistic* [28], equação 2.1, sendo este um ratio que reflete as similaridades dentro do grupo e as diferenças entre grupos.

$$\text{Calinski - Harabasz Pseudo FStatistic} = \frac{\left(\frac{R^2}{n_c - 1}\right)}{\left(\frac{1 - R^2}{n - n_c}\right)} \quad (2.1)$$

Sendo  $n$  o número de *features* e  $n_c$  o número de *clusters*. O valor de  $R^2$ , equação 2.2, é calculado tendo em conta os valores de similaridade intra-*cluster* e a dissimilaridade entre *clusters*.

$$R^2 = \frac{SST - SSE}{SST} \quad (2.2)$$

Neste caso, o SST, equação 2.3, demonstra a diferença entre *clusters*, pois esta tem em conta o valor da variável na *feature* com o valor médio global da variável.

$$SST = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} (V_{ij}^k - \bar{V}^k)^2 \quad (2.3)$$

Sendo  $n_i$  o número de *features* no *cluster*  $i$ ,  $n_v$  o número de variáveis para efetuar o *clustering* das *features*,  $V_{ij}^k$  o valor da  $k$ -ésima variável na  $j$ -ésima *feature* no *cluster*  $i$ , e  $\bar{V}^k$  a média do valor da  $k$ -ésima variável.

O SSE demonstra a similaridade intra-*cluster*, equação 2.4, pois tem em conta o valor da variável da *feature* no *cluster* com o valor médio da variável no *cluster*.

$$SSE = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} (V_{ij}^k - \bar{V}_t^k)^2 \quad (2.4)$$

Sendo  $\bar{V}_t^k$  o valor médio da  $k$ -ésima variável no *cluster*  $i$ .

O número de *clusters* escolhido é aquele que apresentar um maior valor desta estatística.

O pseudocódigo do algoritmo encontra-se na figura 2.14.



```

Escolher centróides iniciais  $c_1, \dots, c_k$ .
...
Enquanto os clusters ainda estão a alterar-se
.. Reatribuir os pontos de dados
.. Para  $i = 1, \dots, n$ 
... Atribuir os pontos de dados  $p_i$  ao cluster
... cujo centro  $c_j$  é o mais próximo
.. terminar
.. Atualizar os centróides dos clusters
.. Para  $j = 1, \dots, k$ 
...  $n_j$  = número de pontos em  $C_j$ ;
...  $c_j = \frac{1}{n_j} \sum_{p_i \in C_j} p_i$ 
.. terminar
terminar

```

Figura 2.14 - Pseudocódigo do Algoritmo K-means ( Adaptado de [30])

O primeiro passo do algoritmo é a escolha dos centróides iniciais, figura 2.15, sendo que o método mais básico é escolher  $n_c$  amostras do conjunto de dados.

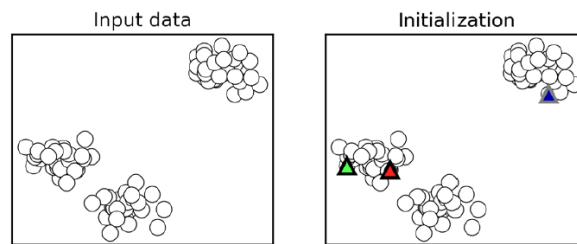


Figura 2.15 - Primeiro Passo do algoritmo "K-means" (Adaptado de [27])

A partir daí o algoritmo consiste em dois passos repetitivos, presentes na figura 2.16. No primeiro as restantes amostras são distribuídas pelos grupos que contêm o centróide mais perto, considerando que  $c_i$  é a coleção de centróides no conjunto de dados  $C$ , então cada ponto de dados  $x$  é atribuído a um *cluster* tendo em conta equação 2.5.

$$\operatorname{argmin} \operatorname{dist}(c_i, x)^2 \quad (2.5)$$

Sendo que *argmin* minimiza os valores da função *dist*, em que esta calcula a distância euclidiana.

Considerando  $S_i$  o conjunto de dados pertencentes ao *cluster*  $i$ , o centróide do grupo é recalculado tendo em conta a média de todos os valores das amostras atribuídas ao centróide anterior, equação 2.6.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i \quad (2.6)$$

O algoritmo termina quando os novos centróides gerados não se alteram de forma significativa.



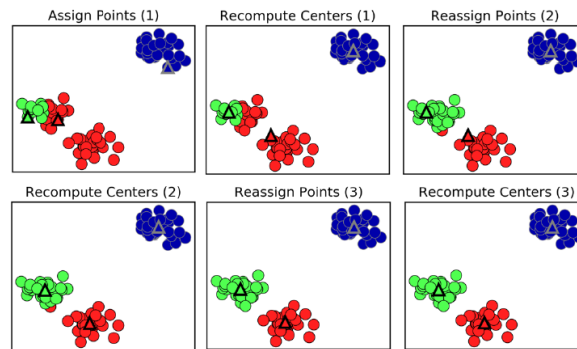


Figura 2.16 - Segundo e Terceiro passo do algoritmo "K-means"(Adaptado de [27])

Na área da saúde, a segmentação tem-se focado muito em atributos individuais e domésticos, como o género, idade, rendimento, composição do agregado familiar, classe social, entre outros. Adicionar atributos como hábitos de vida ou de personalidade, como por exemplo, crenças, traços de personalidade ou até comportamentos, resultaria numa melhor perceção dos subgrupos, resultando em intervenções melhor direcionadas [25].

Agentes de cuidados de saúde já agrupam a população de forma intuitiva com o propósito de prestar esses cuidados de forma efetiva, seja pela sua condição clínica ou por outro tipo de fatores [3]. O agrupamento da população é importante devido a várias razões, tais como:

- Ajuda a entender diferentes necessidades de diferentes partes da população – este é o primeiro passo para alcançar melhores resultados em termos de cuidados integrados, pois a compreensão das características da população faz com que exista escolhas mais adequadas de esquemas e serviços a serem oferecidos.

- Define as combinações principais de cuidados que a população necessita – o principal objetivo dos agentes de cuidados de saúde é lidar com a complexidade das necessidades interdependentes da população. Uma medida única é errada, pois diferentes grupos de pessoas necessitam de diferentes cuidados. Idealmente a segmentação seria a nível individual, mas isso levaria a milhões de segmentos por isso é necessário simplificar. Uma abordagem correta é intuitiva para os profissionais, não se sobrepõe com outros grupos e utiliza atributos sensíveis o suficiente para descrever de forma correta a necessidade dos indivíduos.

- Ajuda a priorizar e implementar uma abordagem por fases – através do agrupamento é possível verificar o que é mais relevante para a população local. Cada área local pode decidir em que grupos pretendem focar-se primeiro, de acordo com prioridades locais e contexto.

Em Inglaterra, o departamento de saúde do Governo efetuou um estudo denominado por “The Healthy Foundations LifeStages Segmentation”, que teve como objetivo principal a criação de uma segmentação utilizando um questionário quantitativo geral da população do seu país. Este estudo foi utilizado para entender melhor os hábitos de vida, principalmente em termos de hábitos alimentares, de consumo de tabaco e álcool. Primeiramente foi projetado para ser utilizado a nível nacional, no entanto, acabou por ser desenvolvido e refinado para ter aplicações locais. Este modelo permite às pessoas identificar e selecionar segmentos locais, sendo que é dirigido a qualquer indivíduo que esteja envolvido no planeamento, desenvolvimento e projeção de marketing social ou de programas de ajuda para alterar comportamentos, seja autoridades locais ou governamentais [25].

Graças à análise de *clusters* foram identificados 5 grupos, presentes na figura 2.17. Estes grupos podem ser encontrados em qualquer estrato social, desde os mais desprivados aos mais abastados.

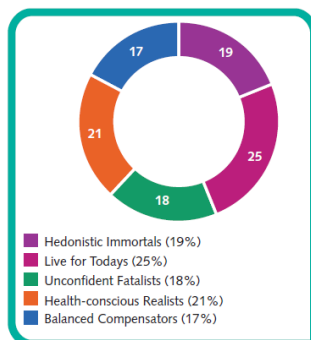


Figura 2.17 - Segmentos do projeto Healthy Foundations. (Extraído de [25])

O grupo “*Hedonistic Immortals*”, figura 2.18, é caracterizado por indivíduos com uma idade média mais baixa, é menos provável que sejam de áreas mais desfavorecidas, são motivados por situações de risco e não demonstram estar preocupados com a sua saúde, sendo que é mais provável que consumam grandes quantidades de álcool e existe uma maior incidência no consumo de drogas.

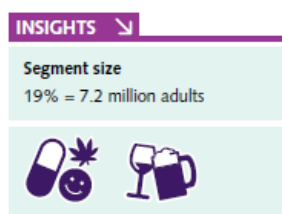


Figura 2.18 - Grupo “*Hedonistic Immortals*” - Healthy Foundations. (Extraído de [25])

Os indivíduos do grupo “*Live for Todays*”, figura 2.19, tem tendência a pertencer a áreas mais desfavorecidas, sendo que demonstram poucos cuidados a nível da sua saúde e pouca preocupação sobre o seu bem-estar no futuro, uma vez que pensam que vão ter pouca longevidade na sua vida. São mais propensos a fumar e a consumir álcool em grandes quantidades.

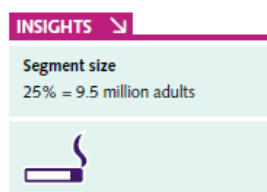


Figura 2.19 - Grupo “*Live for Todays*” - Healthy Foundations. (Extraído de [25])

No grupo “*Unconfident Fatalists*”, figura 2.20, os seus indivíduos têm uma faixa etária mais elevada, vivem em áreas desfavorecidas e tem maior tendência a estarem reformados em vez de empregados. É o grupo que demonstra comportamentos mais negativos em termos de saúde e tem percepções negativas de um estilo de vida saudável.

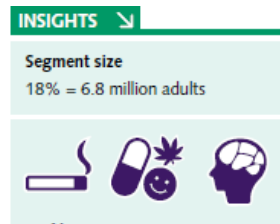


Figura 2.20 - Grupo "Unconfident Fatalists" - Healthy Foundations. (Extraído de [25])

Os indivíduos do grupo “*Health-Conscious Realists*”, figura 2.21, são maioritariamente do género feminino, são mais propensos a viver em áreas menos desfavorecidas e encontram-se acima da idade média. Apresentam comportamentos positivos em termos de saúde, são motivados e sentem-se em controlo da sua vida e saúde, apresentam hábitos alimentares saudáveis, sendo que existe uma baixa incidência de consumo de drogas e álcool.

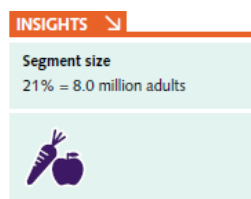


Figura 2.21 - Grupo "Health-Conscious Realists" - Healthy Foundations. (Extraído de [25])

O grupo “*Balanced Compensators*”, figura 2.22, é caracterizado por indivíduos maioritariamente do género masculino, é onde existe um maior número de pessoas empregadas a tempo inteiro, e de forma geral apresentam comportamentos positivos de saúde, praticam exercício físico de forma regular, a sua alimentação é saudável e existe uma baixa prevalência no consumo de tabaco e drogas.

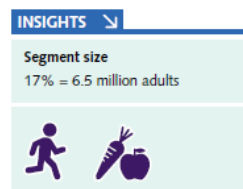


Figura 2.22 - Grupo "Balanced Compensators" - Healthy Foundations. (Extraído de [25])

Este modelo foi utilizado já por algumas entidades, nomeadamente a NHS Tower Hamlets [31] refere que utilizaram este modelo para interligar com outro modelo de segmentação denominado por MOSAIC [32], este mais virado para a vertente de comportamento de consumidor, e queriam entender problemas de saúde que raramente eram resolvidos por intervenções médicas tradicionais e obter informação acerca de hábitos comportamentais para compreender melhor as desigualdades na área da saúde, tendo em conta as suas necessidades e capacidade de mudança.

O programa “Change4Life”[33] tem como objetivo combater as causas de obesidade e utilizou este modelo para entender melhor um novo grupo de interesse que tinham, os adultos entre os 31 e 60 anos. Através desta medida conseguiram entender que pessoas com um peso saudável eram mais propensas a sentirem-se agradas com um estilo de vida saudável e terem alta autoconfiança. As pessoas obesas consideram difícil manter um estilo de vida saudável, tendo baixa autoconfiança e terem um grande índice de fatalidade ao envelhecerem. Estas perceções foram utilizadas para direccionar e desenvolver estratégias de marketing para o novo público alvo.

Variabilidade geográfica na população e a necessidade das pessoas em cuidados de saúde são a base para a análise e planeamento de serviços na área da saúde. A população não se encontra espalhada de forma uniforme no planeta e difere em variadas dimensões, sejam elas idade, género, cultura e economia, que afetam a sua necessidade e possibilidade de recorrer algum tipo de serviço. A necessidade é um conceito multidimensional que reflete as características da população, os seus comportamentos e o ambiente que os rodeia. É neste aspeto que a tecnologia SIG é crucial pois esta faz o mapeamento e explora a variabilidade geográfica na necessidade de serviços de saúde e desenvolver indicadores de necessidade de cuidados de saúde. Tem sido usada durante vários anos para ligar várias camadas de informação sobre a população e o ambiente para caracterizar as variadas necessidades na área da saúde [34].

## 2.5 Análise de Autocorrelação Espacial Local

A análise de autocorrelação espacial local serve para avaliar a significância de estatística local em cada localização da área de estudo, sendo que permite a identificação de *clusters* espaciais e *outliers* espaciais. Um indicador de autocorrelação espacial para uma variável  $y_i$ , observado numa localização  $i$ , com uma estatística  $L_i$  é dado pela expressão 2.7 [4].

$$L_i = f(y_i, y_{j_i}) \quad (2.7)$$

Sendo que  $f$  é uma função e  $y_{j_i}$  são os valores observados na vizinhança  $J_i$  de  $i$ .

A vizinhança  $J_i$  pode ser definida de variadas formas, sendo que esta deverá ser escolhida de acordo com as características dos dados [35]. Existem várias definições de vizinhança espacial que não são exploradas neste projeto, mas a título de exemplo, os métodos de distância inversa são mais apropriados para dados contínuos ou modelos em que quanto mais próximas duas *features* se encontrarem no espaço, mais provável é estas interagirem ou se influenciarem. Ou seja, é um conceito de modelo de relação espacial que tem como base impedências, quanto mais longe uma *feature* se encontra menor será o impacto desta. Também existe a conceptualização de contiguidade de polígonos, neste caso escolhe polígonos vizinhos tem em conta se estes partilham apenas uma aresta, ou também um canto. Este método é efetivo quando os polígonos são similares em forma e distribuição e quando a relação espacial é função de proximidade, ou seja, caso dois polígonos partilhem uma fronteira, a interação espacial entre estes aumenta [36].

Neste projeto o índice utilizado para efetuar esta análise foi o Anselin Local Moran [4]. Este foi concebido como uma ferramenta estatística para testar a autocorrelação local e para detectar objetos espaciais, nomeadamente *clusters* e *outliers*. A equação 2.8 apresenta os cálculos necessários para obter o valor deste índice [35].

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^n w_{i,j} (x_j - \bar{X}) \quad (2.8)$$

Onde  $x_i$  é um atributo da *feature*  $i$ ,  $\bar{X}$  é a média do correspondente atributo,  $w_{i,j}$  é o peso espacial entre a *feature*  $i$  e  $j$ , e  $S_i^2$  é dado pela equação 2.9.

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^n (x_j - \bar{X})^2}{n - 1} \quad (2.9)$$

Onde  $n$  é o número total de *features*.

Valores positivos deste índice revelam que a *feature* tem *features* vizinhas que apresentam similaridade no valor dos atributos, sejam estes elevados ou reduzidos, esta *feature* é parte de um *cluster*. Se o valor for negativo, significa que existe dissimilaridade no valor do atributo da *feature* com as *features* vizinhas, neste caso a *feature* é considerada como *outlier*. Ou seja, graças a este índice é possível verificar zonas geográficas que apresentem similaridade no valor dos atributos, *clusters* espaciais, e zonas geográficas que apresentem dissimilaridade no valor dos atributos, *outliers* espaciais.

Um estudo nos Estados Unidos da América (EUA) [37] quis entender qual o padrão espacial da taxa de natalidade na adolescência e recorreu a este índice para identificar e analisar os *clusters* com elevados valores de taxa de natalidade e os de valor reduzido. Através dos resultados da figura 2.23, os autores conseguiram entender que o padrão espacial pouco se alterou entre 2003 e 2012, sendo que em 2003, 30 condados foram consideradas zonas de “High-Low Outliers”, ou seja, os condados apresentavam elevadas taxas de natalidade, mas estavam rodeados de condados com taxas de natalidade reduzidas. 14 condados foram considerados “Low-High Outliers”, significando que estes apresentavam uma baixa taxa de natalidade, mas estavam rodeados de condados em que este valor era elevado. Em 2012 estes números mantiveram-se relativamente estáveis, existindo cerca de 24 condados como “High-Low Outliers” e 14 como “Low-High Outliers”.

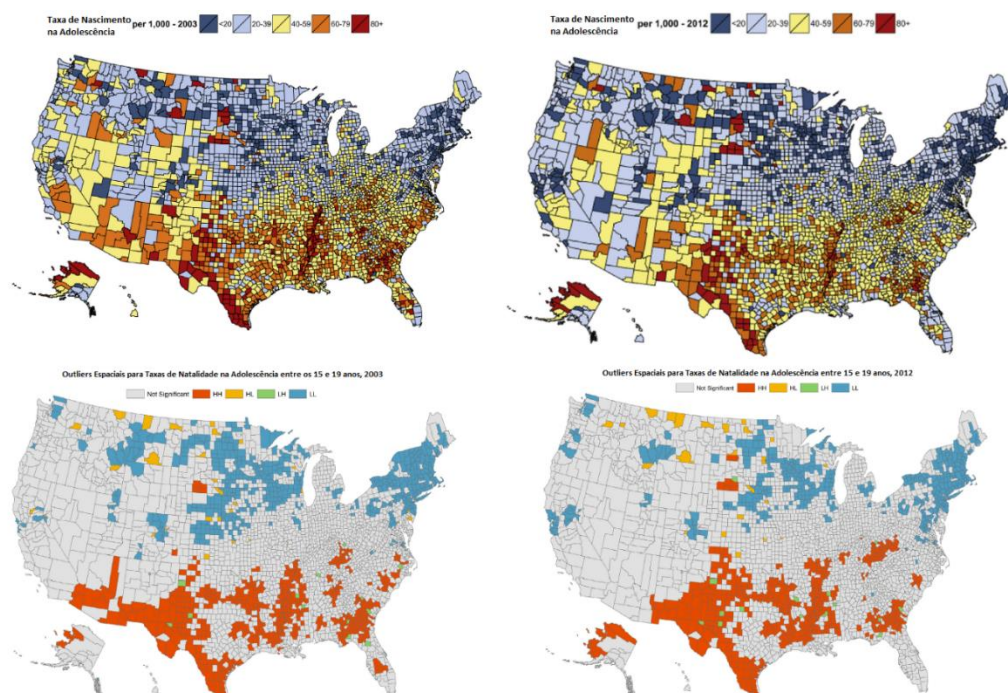


Figura 2.23 - Estimativas baseadas em modelo de Taxa de Natalidade na Adolescência em 2003 e 2012 e respetivos clusters e outliers espaciais (Adaptado de [37])

Os autores concluíram que existe uma variação geográfica substancial nos condados dos EUA, com uma tendência a formarem *clusters* que não era expectável. Estes admitem que esta metodologia pode ajudar a que medidas sejam tomadas no sentido de mitigar e reduzir as taxas de natalidade na adolescência em áreas específicas do país, podendo ser utilizada por agentes de cuidados de saúde ou outros oficiais que sejam responsáveis por informar e desenvolver medidas de prevenção nesta área de modo a facilitar e entender onde e a quem dirigir estas mesmas.

Autores de outro estudo nos EUA [38] quiseram tentar entender qual era o impacto das características da vizinhança das zonas geográficas deste país na taxa de Obesidade, uma vez que esta apresenta padrões espaciais diferentes. Utilizaram três modelos para se basearem no cálculo do índice

de massa corporal (IMC). Um primeiro modelo que era ajustado tendo apenas em conta a idade, raça e género, o segundo modelo incluía adicionalmente dados sócioeconómicos (educação e rendimentos), e o terceiro foi ajustado para incluir variáveis de áreas vizinhas (densidade populacional e dados sócio-económicos). Ao utilizar o índice local de Moran, figura 2.24, no modelo 1 foram apresentados 16 “High-High Clusters” e 44 “Low-Low Clusters” enquanto que no modelo 2 foram apresentados 9 “High-High clusters” e 10 “Low-Low Clusters”, sendo que se verificou uma diminuição no número destes *clusters* ao incluir outro tipo de variáveis. No entanto o modelo 3 não apresentou qualquer tipo de *cluster*.

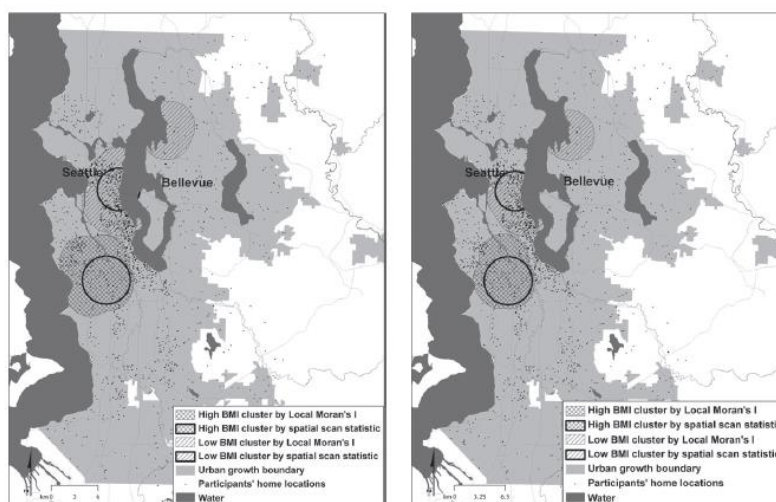


Figura 2.24 - Clusters de valores elevados e reduzidos de índice de massa corporal (Extraído de [38])

Os autores concluíram que a obesidade se concentra espacialmente em *clusters* de valor elevado e reduzido de IMC, sendo que verificaram que zonas mais a Norte tinham como característica populações com um IMC mais reduzido, com maior nível de rendimento, melhor educação e com maior densidade populacional, sendo que em contrapartida, numa zona mais a Sul existe um maior valor de IMC, menor nível de rendimento e uma menor densidade populacional. Compreenderam que a característica da vizinhança poderia explicar as disparidades geográficas sobre a obesidade, uma vez que quando esta era incluída num modelo nenhum *cluster* foi previsto. Estes consideraram uma boa ferramenta para identificar e explicar de forma espacial a obesidade, sendo que poderá servir para definir estratégias de prevenção e intervenção.

### 3. Segmentação

Este capítulo contém toda a informação relativa à criação da segmentação. Assim, no primeiro subcapítulo descrevem-se os dados e as suas principais características. Os dois restantes subcapítulos são referentes à criação da segmentação e respetivos resultados, assim como toda a análise de autocorrelação espacial obtida através da aplicação do índice local de Moran, terminando com uma análise a ambos os resultados. Será primeiramente feita uma análise exploratória de áreas à distribuição espacial dos atributos e respetiva análise ao índice local de Moran. Por fim tecem-se algumas considerações sobre os resultados de ambas as metodologias e demonstra-se como ambas podem atuar complementarmente para a caracterização da população na área da saúde. Serão discutidos em detalhe alguns dos resultados obtidos e interligar os resultados da segmentação e do índice local de Moran.

#### 3.1 Metadados

Os dados sociodemográficos utilizados foram os principais dados alfanuméricos e geográficos acerca dos censos de 2011. Estes estão inseridos numa camada de informação denominada por Base Geográfica de Referenciação de Informação (BGRI), que é um ficheiro que contém polígonos, e o seu sistema de coordenadas é o *ETRS 1989 TM06-Portugal*. O detalhe mais fino deste ficheiro é a subsecção estatística.

A tecnologia ESRI disponibiliza uma ferramenta geoprocessamento denominada de *Enrich Layer*, sendo que esta adiciona dados demográficos sobre as pessoas ou lugares que rodeiam ou estão presentes nas camadas de informação [39]. A fonte de dados desta ferramenta é a consultora Michael Bauer Research [40]. Esta é uma ferramenta que permite enriquecer camadas de informação geográfica com variados atributos tendo em conta a localização destes, caso os dados se encontrem disponíveis ao nível geográfico pretendido. Esta base de dados tem um conjunto de indicadores compilados em 5 grupos: População, Dados sociodemográficos alargados (Agregados familiares, Género e Faixa etária, Educação, Estado civil, Empregabilidade, etc), Número de famílias por rendimento, Poder de compra e Dados de gastos dos consumidores por grupos de produtos [40].

Os dados selecionados para complementarem os dados provenientes da BGRI e que provieram desta ferramenta foram os seguintes: Gasto em álcool (gasto em euros, são aquelas que são compradas para consumo doméstico. Não são tidas em conta as bebidas consumidas em hotéis, restaurantes, cafés, bares, entre outros), Gasto em tabaco (gasto em euros, são tidas em conta todas as compras de tabaco seja em restaurantes, bares, cafés, estações de serviço, entre outros e famílias por nível de rendimento. A nível nacional, o número de famílias encontra-se distribuídos em 5 classes de acordo com o seu rendimento. No 1º quintil encontram-se famílias com um rendimento inferior a 13,296 €, no 2º quintil encontram-se famílias com um rendimento entre os 13,296 € e 19,250 €, no 3º quintil encontram-se famílias com um rendimento entre os 19,251 € e 26,418€, no 4º quintil encontram-se famílias com um rendimento entre os 26,419€ e 38,608€, sendo que no 5º quintil encontram-se as famílias com um rendimento superior aos 38,609€.

#### 3.2 Metodologia

Neste subcapítulo são apresentadas as ferramentas que foram utilizadas para a criação da segmentação e para a análise espacial realizada posteriormente assim como são descritos os procedimentos necessários para a sua aplicação.



### 3.2.1 Seleção das variáveis relevantes

Os dados selecionados têm de ser relevantes para a área da saúde, e principalmente na área da diabetes. Variada bibliografia afirma que a diabetes é uma doença que é afetada pelas condições socioeconómicas da população. Por exemplo, os autores de um estudo [20] utilizaram variáveis como rendimento, educação e empregabilidade para tentar entender quais as condições e características da população diabética que se encontrava controlada. Outro estudo [41] concluiu que lugares e comunidades envelhecidas, com baixa escolaridade, desemprego elevado, baixos rendimentos, considerando estes grupos socioeconómicos mais desfavorecidos são ambientes vulneráveis e determinantes para a incidência da diabetes. A hipertensão arterial e a diabetes são doenças inter-relacionadas. Normalmente no momento de diagnóstico da diabetes, a hipertensão já existe em cerca de 40% das pessoas, o que sugere uma relação entre as duas [42]. Existem fatores de risco que é necessário ter em atenção, nomeadamente o tabagismo e alcoolismo [42].

Neste sentido, foram utilizados dados que caracterizassem a população, sendo que os poucos disponíveis publicamente e que fariam sentido utilizar para a área da saúde estão presentes nos dados dos censos de 2011. No caso da BGRI, foram apenas selecionados os dados que caracterizassem o indivíduo, excluindo todos os atributos relativos aos edifícios e à constituição do agregado familiar. Para complementar esta informação foram ainda utilizados mais atributos provenientes da base de dados da ESRI, já referida anteriormente. Esta tem uma panóplia de dados de gastos dos consumidores, como por exemplo gasto em comida e bebidas não-alcoólicas, gasto em roupa, entre outros. O gasto em comida e bebidas não-alcoólicas seria um atributo interessante a considerar caso discriminasse o tipo de alimentos. Assim poderíamos inferir, por exemplo, que onde existisse um maior gasto em alimentação saudável seria onde os indivíduos teriam uma maior preocupação com a sua alimentação. É com este tipo de raciocínio que foram selecionados dois atributos: gasto em álcool e gasto em tabaco. De certa forma ao termos acesso onde este gasto é maior, podemos presumir que a população tem uma maior tendência em consumir este tipo de substâncias que são nefastas para a sua saúde.

Os dados utilizados para fazer a segmentação e a respetiva análise espacial de *cluster* foram os seguintes: BGRI (INE), Famílias por nível de Rendimento (ESRI), Gasto em álcool (ESRI) e Gasto em tabaco (ESRI). No total foram utilizados cerca de 27 atributos, estando estes divididos em 7 categorias: **Género** (número de Indivíduos residentes do género Masculino e género Feminino), **Faixa etária** (número de indivíduos residentes com faixa etária compreendida entre 0 e 4 anos, 5 e 9 anos, 10 e 14 anos, 15 e 19 anos, 20 e 24 anos, 25 e 64 anos e 65 ou mais anos de idade), **Empregabilidade** (número de indivíduos residentes desempregados à procura do 1º emprego, desempregados à procura de emprego, empregados, pensionistas e reformados e sem atividade económica), **Ensino** (número de indivíduos residentes com o 1º ciclo, 2º ciclo, 3º ciclo, ensino secundário concluído, ensino pós-secundário e com o ensino superior concluído), **Rendimentos** (número de famílias ao 1º Quintil, 2º Quintil, 3º Quintil, 4º Quintil e 5º Quintil), **Gasto em tabaco** e **Gasto em álcool**.

### 3.2.2 Modelo Utilizado

Neste projeto, o objetivo da segmentação é conseguir segmentar a população ao nível das freguesias de acordo com as suas características internas e não comparar a nível nacional. Para esse efeito, o método adotado foi converter as suas características em percentagem. O objetivo seria, por exemplo, responder a esta questão “Nesta freguesia existe maior percentagem de pessoas desempregadas, empregadas, pensionistas ou sem atividade económica?”. Uma vez que através da BGRI temos acesso ao número total de residentes, cada atributo proveniente da camada de informação “BGRI” foi dividido pelo número total de residentes, de modo a mostrar qual a percentagem desse atributo em



comparação aos outros nessa mesma categoria. No caso dos dados provenientes da ESRI, esta medida apenas foi adotada nos atributos referentes a “Famílias por nível de rendimento”, no qual se criou um atributo que fosse a soma do número de todas as famílias em todos os quintis e em seguida dividir os atributos por esse total.

Para este efeito foi necessário criar uma camada de informação que contivesse todos estes atributos. A resolução espacial escolhida foi ao nível da freguesia. O primeiro passo foi efetuar a ferramenta de geoprocessamento *Dissolve*, na camada de informação “BGRI”, pois a sua resolução espacial é a subsecção estatística, sendo que o *Dissolve Field* é o campo que contém o código DICOFRE, e apenas foram escolhidos 20 atributos, relativos ao indivíduo, pertencentes às seguintes categorias: Género, Faixa Etária, Empregabilidade, Ensino. A estatística aplicada aos campos foi a soma. Este processo gerou uma camada de informação denominada “BGRI\_Freguesias”, sendo que esta era uma camada de polígonos com todos os atributos desejados a nível de freguesias provenientes do ficheiro “BGRI”.

Para enriquecer esta camada de informação com dados provenientes da base de dados da ESRI recorreu-se à ferramenta de geoprocessamento *Enrich Layer*. Como referido no subcapítulo 3.1, esta é uma ferramenta que enriquece camadas de informação com dados sociodemográficos sobre a população que se encontra inserida nessa mesma camada. Neste caso, uma vez que a camada de informação contém todas as freguesias de Portugal continental, esta ferramenta irá criar uma nova camada de informação a nível de freguesias com a adição das variáveis selecionadas sobre a população das freguesias. Como *Input Features* foi selecionada a camada de informação “BGRI\_Freguesias”, sendo esta a camada que foi enriquecida, e como *Variables* foram escolhidos os seguintes 7 atributos: Número de famílias no 1º Quintil, 2º Quintil, 3º Quintil, 4º Quintil e 5º Quintil de Rendimentos, Gasto em Tabaco e Gasto em Alcool.

Concluído este processo, a camada de informação “Informação\_Freguesias” foi gerada, sendo esta uma camada de informação de polígonos com todos os dados necessários para efetuar a segmentação e a respetiva análise espacial de *clusters*. O mapa com a distribuição espacial de cada atributo encontra-se no Anexo I.

Para a criação da segmentação, o passo seguinte foi utilizar a ferramenta *Multivariate Clustering*. Esta é uma ferramenta que encontra *clusters* baseados apenas no valor dos atributos. Esta após introduzido o número de *clusters* desejado, irá procurar por uma solução na qual todas as *features* inseridas num *cluster* são as mais similares entre si e todos os *clusters* são o mais diferentes entre si possível. Como parâmetro de escolha é utilizado o *Calinski-Harabasz pseudo F-statistic*. A similaridade de atributos é baseada no conjunto de atributos que é especificado, sendo que os *clusters* são criados através do algoritmo *K-means*. Neste caso, para a ferramenta assegurar que a solução escolhida é aquela que maximiza a similaridade intra-*cluster* e a diferença entre *clusters* esta vai tentar todas as combinações possíveis da combinação dos atributos. Neste caso a ferramenta apenas avalia as soluções de 2 a 30 *clusters*.

Como *Input Features* foi selecionada a camada de informação “Informação\_Freguesias”, como *Clustering Method* foi selecionado o “K means”, como *Initialization Method* foi utilizado o “Optimized seed locations”, pois este é o método que vai otimizar os resultados e o desempenho da análise. Este método primeiramente escolhe a seed, ou seja, o centroide inicial, de forma aleatória, mas assegura que os centróides subsequentes selecionados representam *features* que se encontram distantes entre si no espaço de dados. O *Number of Clusters* primeiramente foi deixado vazio, o que fez com que a ferramenta avaliasse o número ideal de *clusters* ao computar um Pseudo F-Statistic para soluções de 2 a 30 *clusters*.

Nesta avaliação o número sugerido de *clusters* foi de 2, como se pode ver no gráfico 3.1, pois o critério utilizado neste caso é aquele que apresenta um valor maior de estatística pseudo-F. Este valor indica as soluções que têm um melhor desempenho a maximizar a similaridade dentro do *cluster* e diferenças entre *clusters*.

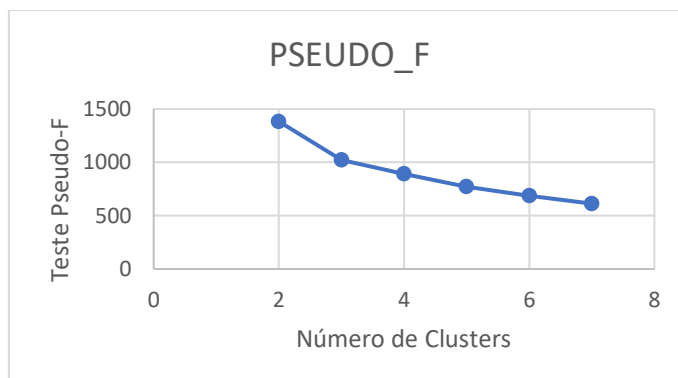


Gráfico 3.1 - Estatística Pseudo-F do número de Clusters

Este é o critério utilizado quando não existe preferência na quantidade de *clusters* que deva existir [28]. Neste projeto testou-se três soluções: apenas dois *clusters*, três *clusters* e quatro *clusters*, presentes na figura 3.1.

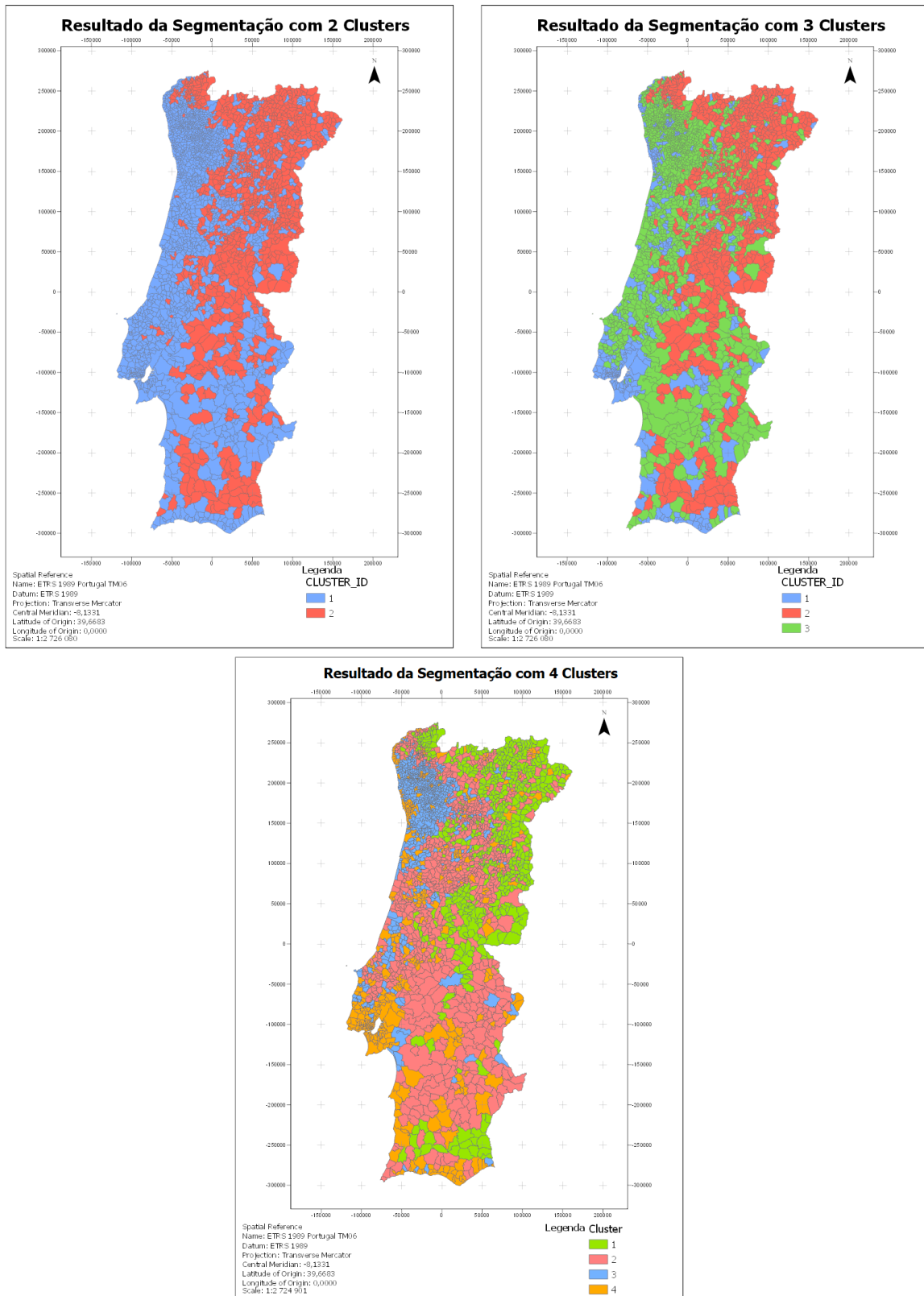


Figura 3.1 - Três soluções de segmentação geradas com 2, 3 e 4 clusters

A solução de 2 clusters não acrescenta grande informação e impossibilitaria de fazer algo para além de uma análise superficial, uma vez que apenas existem dois grupos diferentes, um no qual a população teria um nível mais alto de escolaridade, melhores rendimentos e uma faixa etária mais baixa.

A solução com 3 *clusters* segmenta a população por freguesias com maior percentagem de pessoas adultas com rendimentos altos, nível de estudos superior, freguesias com valores de percentagem com valores médios nos atributos, e freguesias com valores mais elevados na população idosa e pensionistas e reformados. O importante neste caso é que se consiga diferenciar os grupos tendo sido adotada a solução com 4 *clusters*. No final deste processo foi originada a camada de informação “Segmentação”, sendo esta uma camada de polígonos com toda a informação referente à segmentação, onde cada polígono tem associado o número do *cluster* que lhe foi atribuído.

### 3.2.3 Autocorrelação Espacial Local

O objetivo desta análise é compreender com maior rigor como é que os atributos usados no processo de segmentação se distribuem ou se agrupam no espaço e tentar usar essa informação para compreender os resultados da segmentação.

Para efetuar uma análise de autocorrelação espacial dos atributos em estudo foi aplicado o índice de Moran Local que se encontra na ferramenta de geoprocessamento *Cluster and Outlier Analysis* (*Anselin Local Moran's I*). Esta ferramenta vai gerar um mapa de regiões (áreas representativas de freguesias) classificadas por 5 classes : “High-High Cluster” que são freguesias com o valor do atributo alto e que têm freguesias circunjacentes com valores similares altos; “Low-Low Cluster” que são freguesias com o valor do atributo baixo e que têm freguesias circunjacentes com valores similares baixos; “High-Low Outliers” que são freguesias com o valor do atributo alto que são rodeadas por freguesias com valores dissimilares, ou seja, baixos, e os “Low-High Outliers”, o inverso (“não significantes”). Cada freguesia tem um valor de z-score e um p-value associado, que representam medidas de significância estatística quando é aplicado um teste estatístico de aceitação ou rejeição de hipótese de aleatoriedade espacial na avaliação da autocorrelação espacial dos atributos [35].

Como *Input Feature Class* foi escolhida a camada de informação “Informação\_Freguesias”, sendo que como *Input Field* foram selecionados um a um, os 27 atributos desta camada. Como *Conceptualization of Spatial Relationships* foi escolhido o “Inverse Distance”, pois é o mais apropriado a utilizar quando queremos que quanto mais próximo duas *features* estiverem, mais provável é estas interagirem e terem influência entre si. Como *Distance Method* foi escolhido o “Euclidean” e como *Number of Permutations* foi utilizado “9999”. As permutações são utilizadas para determinar o quão provável é encontrar o padrão espacial dos dados a serem analisados ao comparar esses mesmos valores com um grupo de valores gerados de forma aleatória, por isso é uma boa prática usar o número máximo de permutações possível para fazer esta comparação [35]. Na figura 3.2 encontra-se um exemplo do *output* desta ferramenta e no Anexo I o resultado de todos os atributos.

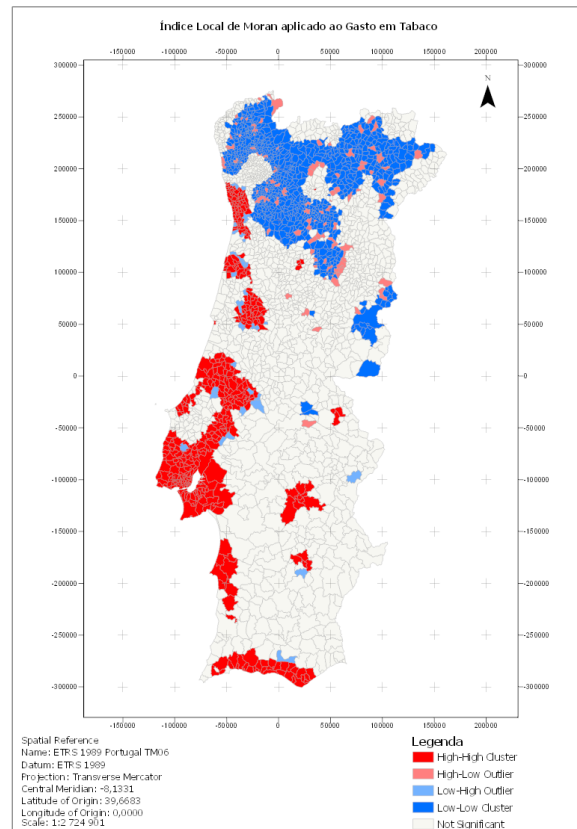


Figura 3.2 - Exemplo do Output do Índice Local de Moran

Como se pode verificar na figura 3.2, existe uma forte presença de “Low-Low Clusters” na zona Norte do continente, ou seja, são freguesias que apresentam valores reduzidos no atributo “Gasto em Tabaco” e as freguesias vizinhas apresentam valores similares. No entanto numa zona mais a litoral no Centro e Sul do continente já encontramos a presença de “High-High Clusters” sendo que esta são freguesias que apresentam valores elevados nesse mesmo atributo e as freguesias vizinhas apresentam valores similares. Com esta análise podemos afirmar que numa zona interior Norte do continente são onde se encontram as freguesias onde existe menor gasto em Tabaco, enquanto que numa zona litoral no Centro e Sul do continente encontram-se as freguesias onde esse mesmo gasto é elevado.

### 3.3 Análise da segmentação

As 2882 freguesias de Portugal continental foram segmentadas da seguinte forma: 603 pertencem ao *Cluster* 1, 1052 ao *Cluster* 2, 755 ao *Cluster* 3 e 472 ao *Cluster* 4, como se pode ver no gráfico 3.2.

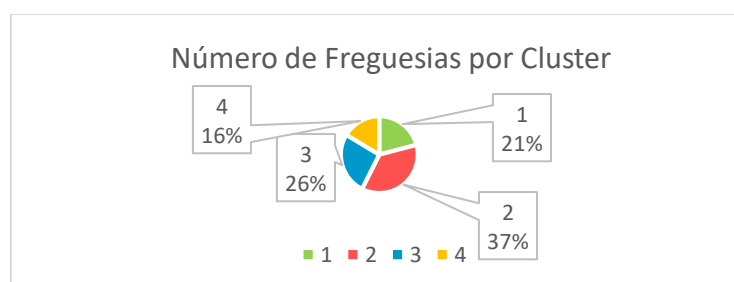


Gráfico 3.2 - Número de freguesias por cluster

O resultado desta ferramenta encontra-se representado na figura 3.3. De certa forma é possível visualizar algum padrão espacial: as freguesias pertencentes ao *cluster* 1, cor verde, encontram-se maioritariamente numa zona interior do Centro-Norte, enquanto que as freguesias pertencentes ao *cluster* 2, cor vermelha, encontram-se mais numa zona Centro-Sul, mas mais no interior. Existe uma grande concentração de freguesias pertencentes ao *cluster* 3, com cor azul, na zona Litoral-Norte, enquanto que o *cluster* 4, cor laranja, encontra-se mais numa zona litoral do Centro-Sul, apesar de existirem umas freguesias na zona litoral Norte.

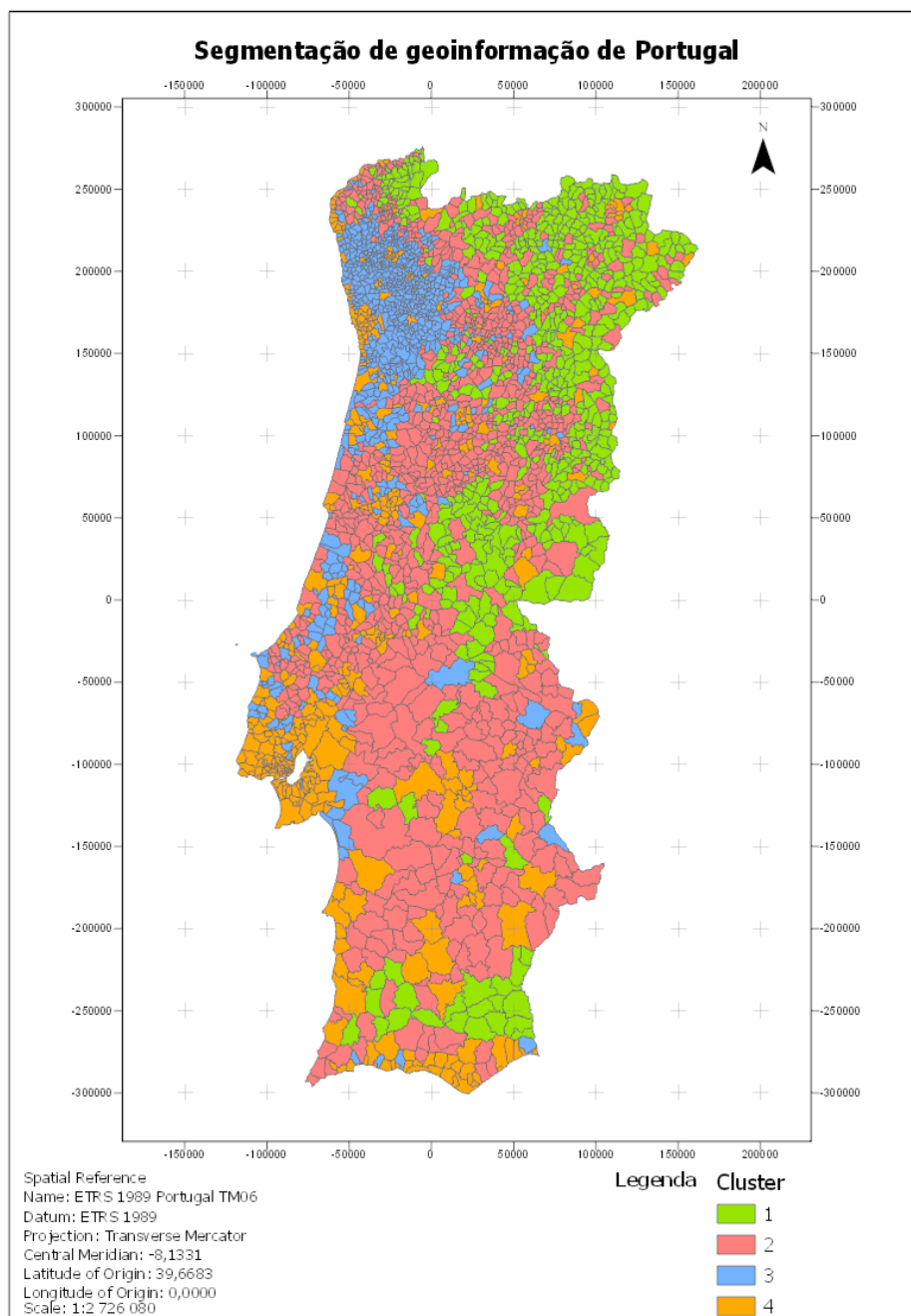


Figura 3.3 - Mapa com o resultado da Segmentação de Geoinformação em Portugal

### 3.3.1 Análise Exploratória dos Atributos

Para a criação da segmentação foram utilizados cerca de 20 atributos, sendo que estes se encontram agrupados em 8 categorias: Género, Faixa etária, Empregabilidade, Ensino, Rendimento, Densidade Populacional, Gasto em Tabaco e Gasto em Álcool.

#### Género

Considerando que os dados representados são valores em percentagem presente em cada freguesia na figura 3.4, quanto mais escura for a cor simbolizada maior será a percentagem. Podemos verificar que os dois tipos de género estão distribuídos de forma diferente espacialmente, pois existe uma maior percentagem de indivíduos do género Masculino nas freguesias do Sul do continente e as freguesias mais no interior Centro e Norte litoral. Esta análise pode ser reforçada através dos mapas com os índices locais de Moran dos respetivos atributos, pois o que importa interpretar são as áreas que representam os “High-High Clusters”, áreas que têm valores altos e a sua vizinhança têm valores similares altos, representados a vermelho forte, e “Low-Low Cluster”, áreas que têm valores baixos e a sua vizinhança têm valores similares baixos, representadas a azul forte. Em termos de género Masculino podemos verificar que os “High-High Cluster” encontram-se grande parte na zona Sul do país, e alguns na zona Norte litoral e interior, sendo que estas mesmas áreas correspondem às áreas com maior percentagem no mapa com a distribuição espacial do atributo, o que seria de esperar. Sobre os indivíduos do género Feminino, verifica-se uma grande concentração na zona Norte Litoral, mesmo na zona onde é feita a fronteira com Espanha. Como seria expectável, estes mapas são complementares.

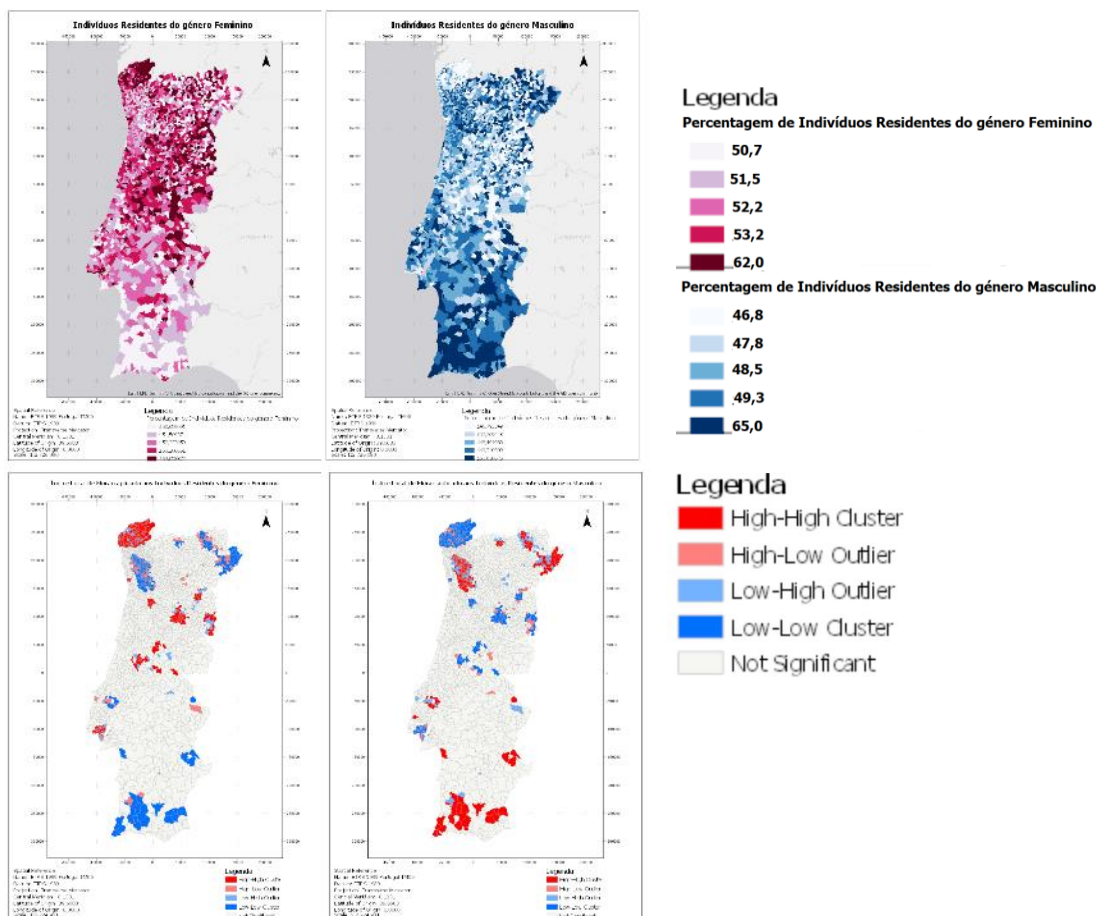


Figura 3.4 - Mapas dos atributos referentes ao Género (à esquerda género Feminino, à direita Masculino) e respetivos índices locais de Moran

Em termos de atribuição aos *clusters*, em Portugal não existe uma grande diferença em termos de género. Por norma existem mais indivíduos do género feminino como se pode ver através do gráfico 3.3, e uma vez que são apenas dois atributos dentro desta categoria, o *cluster* que apresentar uma maior percentagem de indivíduos do género feminino, neste caso o *cluster* 1, é aquele que apresenta menor percentagem de indivíduos do género masculino. Em parte, este gráfico corrobora aquilo demonstrado pela figura 3.4 (página 34), nomeadamente pelo índice de Moran, pois o *cluster* 3 é aquele que apresenta uma maior percentagem de indivíduos do género Masculino. Se olharmos para o mapa da segmentação, figura 3.3 (página 33), verificamos que a zona onde a concentração de freguesias pertencentes ao *cluster* 3 é maior, no distrito do Porto e Vila Real, é onde se encontram também os “High-High Cluster”, demonstrando que nessa freguesia se encontram valores altos desse mesmo atributo e nas freguesias circunjacentes.

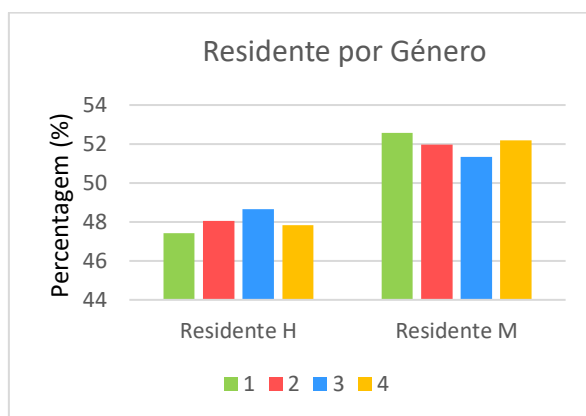


Gráfico 3.3 - Atributos referentes ao Género por Cluster

### Faixa Etária

Em termos de faixa etária, é feita uma divisão nesta categoria. Numa primeira fase é analisada uma faixa etária mais jovem e adolescente, compreendo as faixas etárias de 0 a 4, 5 a 9, 10 a 13 e 14 a 19 anos, sendo que a outra divisão compreende faixas etárias mais adultas de, 20 a 24, 24 a 64 e 65 ou mais anos.

Quanto às faixas etárias de jovens e adolescentes, podemos visualizar na figura 3.5 que as freguesias que têm uma maior percentagem destes indivíduos encontram-se mais na zona litoral do continente, e mesmo em termos de “High-High Clusters”, sendo que se verifica que nos distritos de Porto, Braga são aqueles onde existe maioritariamente os “High-High Cluster”. No distrito de Lisboa verifica-se alguns desses mesmos *clusters*, mas numa faixa etária mais reduzida, entre 0 a 4 anos e 5 a 9 anos. Na zona interior do país, e principalmente a Norte, existe uma forte presença de “Low-Low Cluster” o que significa que as freguesias presentes nesse tipo de *cluster* são freguesias que têm valores baixos de percentagem de indivíduos nesta faixa etária e as freguesias vizinhas também, portanto podemos afirmar que na zona interior mais a Norte do continente existem poucos indivíduos com uma faixa etária entre 0 a 14 anos comparativamente ao resto do país.



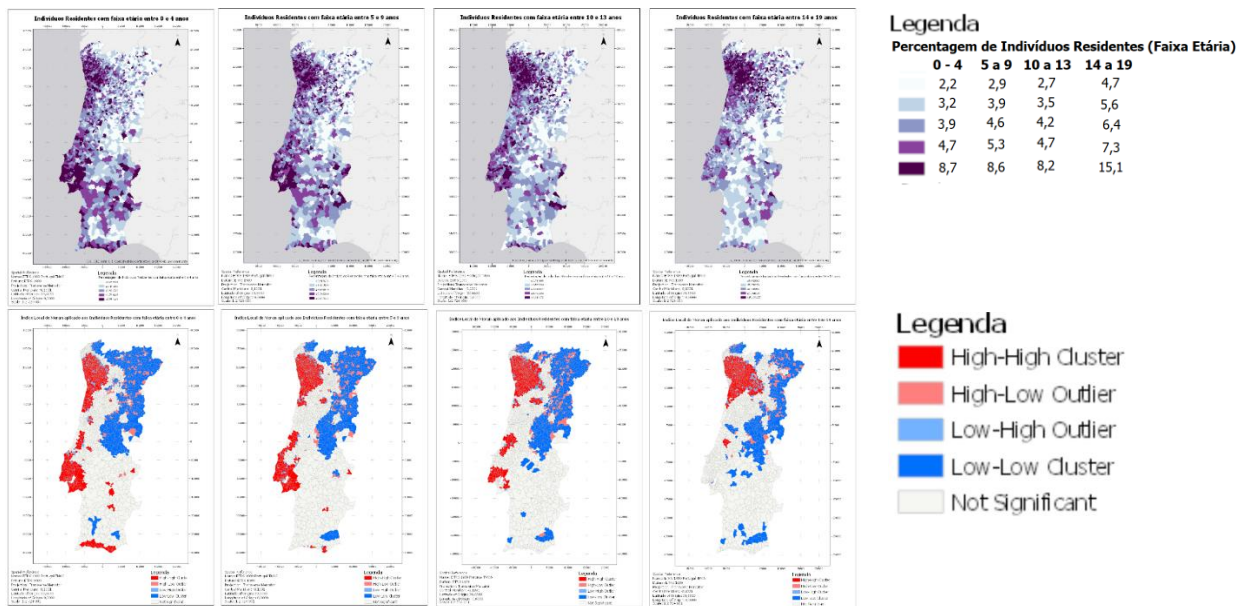


Figura 3.5 - Mapas dos atributos referentes a Faixas etárias de Jovens de Adolescentes (da esquerda para a direita: 0 a 4, 5 a 9, 10 a 13 e 14 a 19) e respetivos índices locais de Moran

Quanto a indivíduos com faixas etárias mais elevadas, figura 3.6, nomeadamente já considerados adultos ou idosos, podemos verificar que quando analisamos os indivíduos residentes com 65 ou mais anos, estes apresentam uma distribuição espacial completamente oposta à distribuição das faixas etárias mais baixas, pois estas normalmente encontram-se localizadas mais numa região litoral enquanto que a faixa etária mais idosa se encontra em zonas de interior, principalmente na zona Norte do país. As faixas etárias de 19 a 24 anos encontram-se situadas em freguesias do distrito do Porto e Braga, na zona Norte do país, sendo que a população adulta, normalmente trabalhadora, a faixa etária de 24 a 64 anos encontra-se mais espalhada numa zona litoral de Centro a Norte do país.

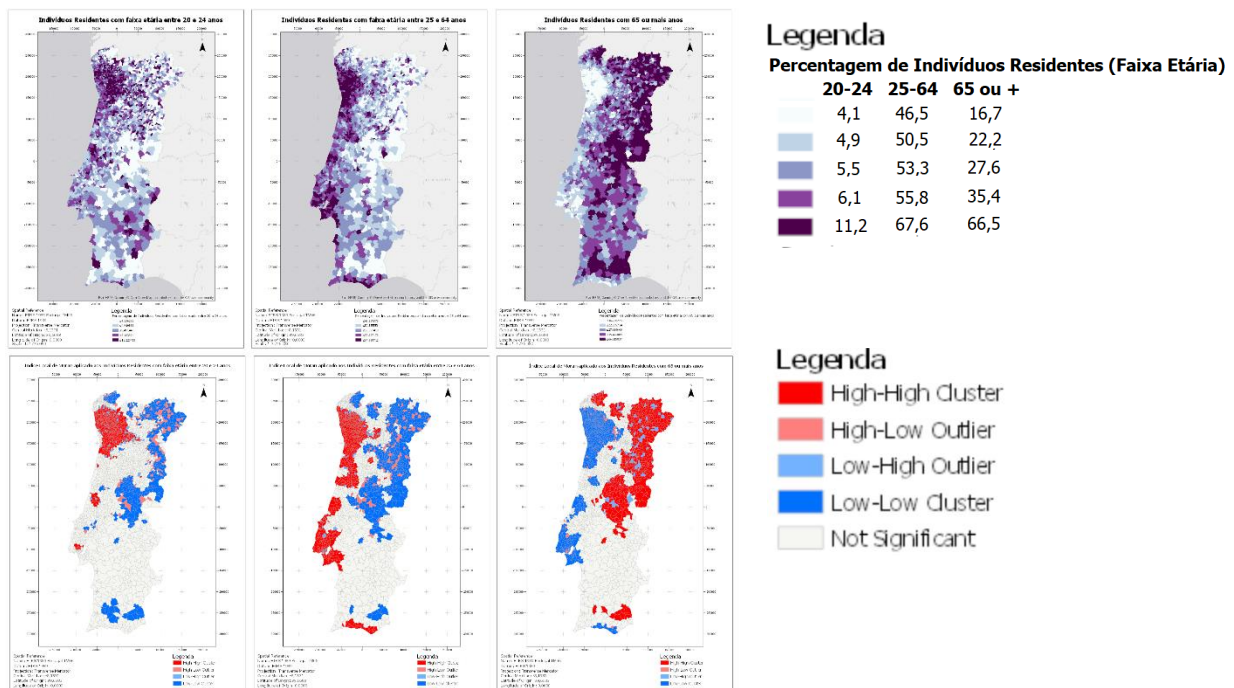


Figura 3.6 - Mapas dos atributos referentes a Faixas Etárias de Adultos e Idosos (da esquerda para a direita: 20 a 24, 25 a 65 e mais de 65 anos) e respetivos índices locais de Moran

Nas faixas etárias mais jovens, gráfico 3.4, podemos verificar que existe uma grande predominância do *cluster* 3 nesta categoria, ou seja, é onde estão presentes as freguesias que têm uma maior percentagem de jovens e adolescentes, e se verificarmos os índices de Moran destes 5 atributos verificamos que os “High-High Cluster” se encontram presentes nas freguesias em que estão presentes os *clusters* 3 e 4, que são aqueles que apresentam uma maior percentagem de indivíduos residentes com estas faixas etárias.

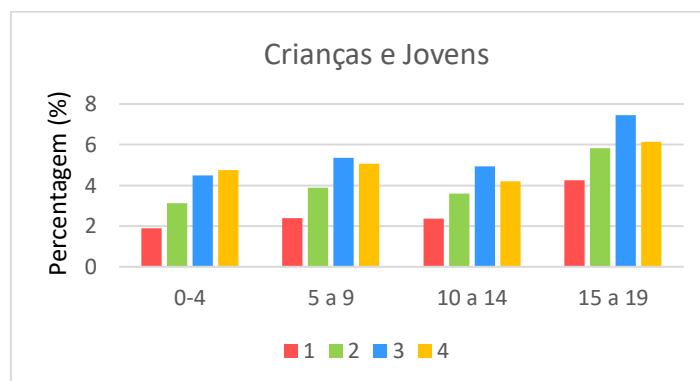


Gráfico 3.4 - Atributos referentes ao Faixa etária de Jovens e Adolescentes por cluster

Ao considerarmos faixas etárias mais altas, nomeadamente adultos e idosos, presente no gráfico 3.5, vemos uma clara predominância do *cluster* 1 na faixa etária de 65 ou mais anos, sendo que se verificarmos o mapa do índice de Moran deste atributo, figura 3.6, as áreas em que se encontram presentes os “High-High Cluster”, mais numa zona interior Norte, é também onde se encontra a maior concentração de freguesias pertencentes a este mesmo *cluster*.

Na faixa etária entre os 25 e os 64 anos, já se verifica uma subida do *cluster* 4 em relação ao *cluster* 3, em comparação às faixas etárias anteriores em que normalmente o *cluster* 3 seria o predominante. Nesta categoria, *cluster* 2 não sobressai em nenhum dos atributos.

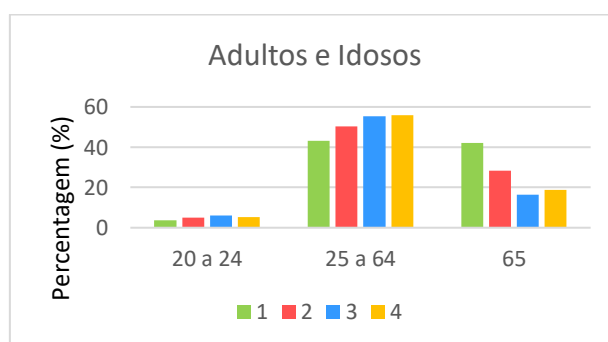


Gráfico 3.5 - Atributos referentes a Faixas Etárias de Adultos e Idosos por cluster

### Ensino

Na área do ensino foram elaborados dois grupos: as pessoas que têm o 1º, 2º e 3º ciclo completo, figura 3.7, e os indivíduos residentes com o ensino secundário, pós-secundário e superior completo, figura 3.8.

Neste primeiro caso, na zona interior Norte verifica-se uma maior concentração de freguesias com grandes percentagens de indivíduos com apenas o 1º ciclo completo, apesar de alguns *outliers* “Low-High”, o que significa que existe algumas freguesias em que esta percentagem é reduzida. Os

indivíduos apenas com o 2º ciclo completo encontram-se maioritariamente concentrados nas freguesias da zona litoral Norte. Em relação aos indivíduos com apenas o 3º ciclo completo, estes já se encontram distribuídos ao longo do país, mas apenas na zona litoral. É a partir deste grau de ensino que a zona do Algarve já começa a ter alguns *clusters* “High-High”.

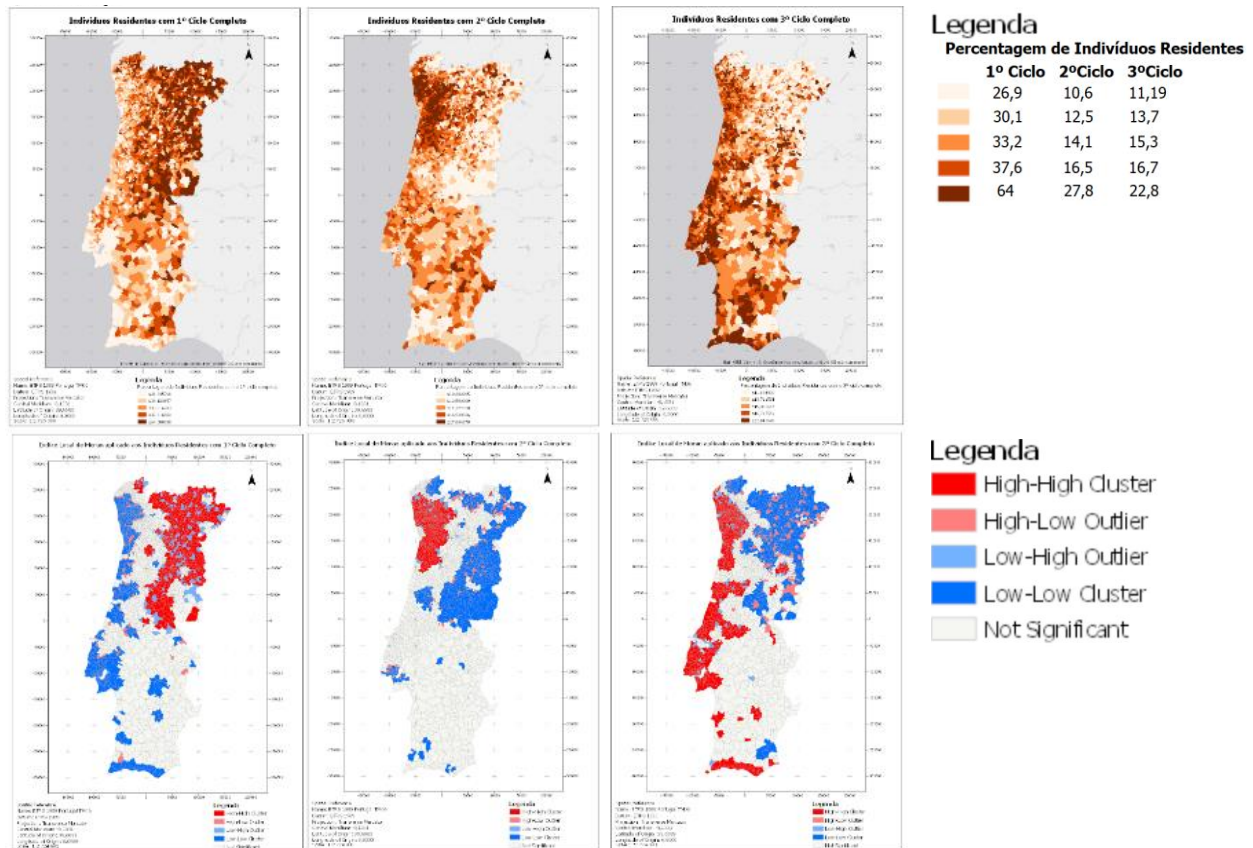


Figura 3.7 - Mapas dos atributos referentes ao Ensino (da esquerda para a direita: 1º, 2º e 3º ciclo) e respetivos índices locais de Moran

Verificando a figura 3.8, no caso do ensino secundário e pós-secundário, os indivíduos com este grau completo encontram-se distribuídos de forma semelhante, mais numa zona litoral ao longo do continente, enquanto que aqueles que têm o ensino superior completo, se encontram maioritariamente nas freguesias presentes no distrito de Aveiro, Coimbra, Lisboa, Castelo Branco, Guarda, Portalegre e Beja. Isto não significa que não possuam freguesias com grande percentagem de indivíduos com o ensino superior completo, pois apenas o distrito de Portalegre é o que não apresenta um “High-Low outlier”, ou seja, é o único distrito em que nenhuma das freguesias apresenta um valor elevado de indivíduos residentes que tenham o ensino superior completo.

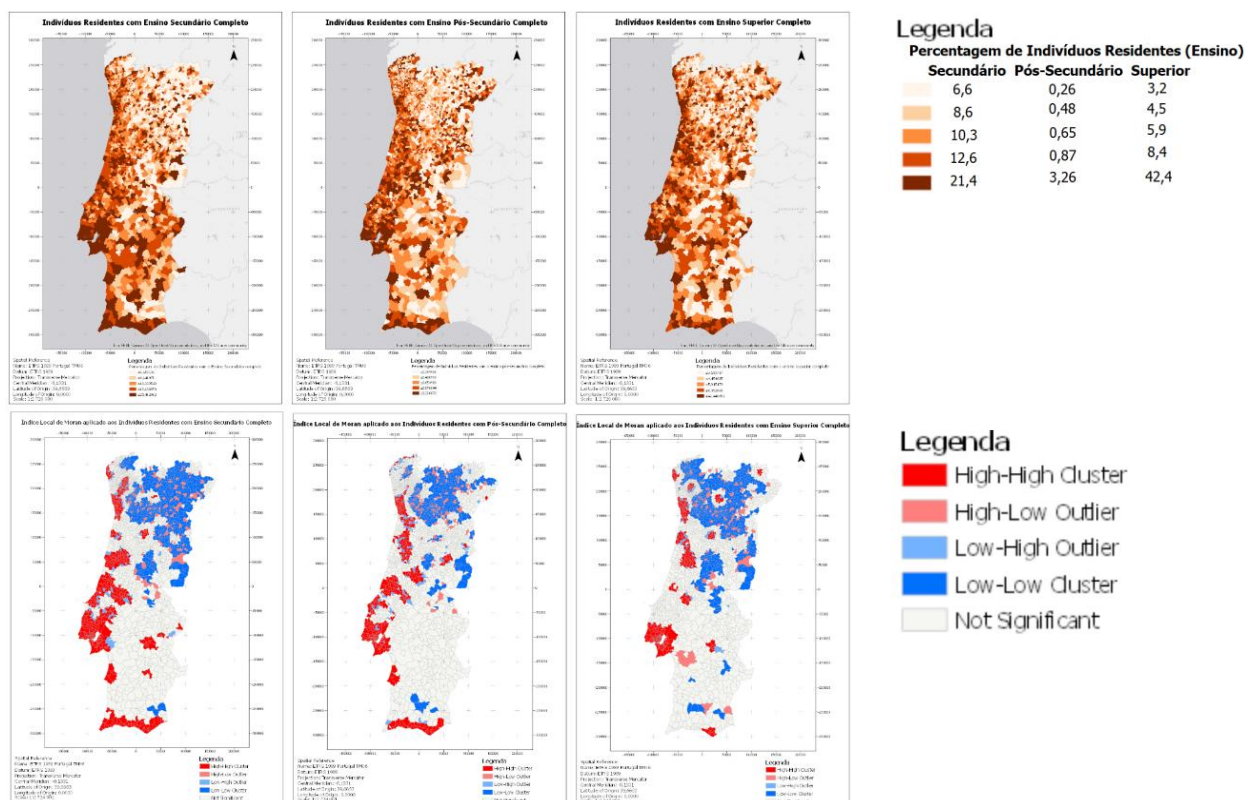


Figura 3.8 - Mapas dos atributos referentes ao Ensino ((da esquerda para a direita: Secundário, Pós-Secundário e Superior) e respetivos índices locais de Moran

Através do gráfico 3.6 podemos afirmar que o *cluster 1* é aquele em que predomina os indivíduos com o 1º ciclo completo, enquanto que o *cluster 2* também tem maiores percentagens nos indivíduos com o 1º e 2º ciclo completo. O *cluster 3* tem maiores percentagem nos primeiros ciclos, o que acaba por corroborar o fato de ter também as maiores percentagens em faixas etárias mais baixas. Os indivíduos com ensino secundário e superior completo encontram-se em maior percentagem nas freguesias presentes no *cluster 4*.

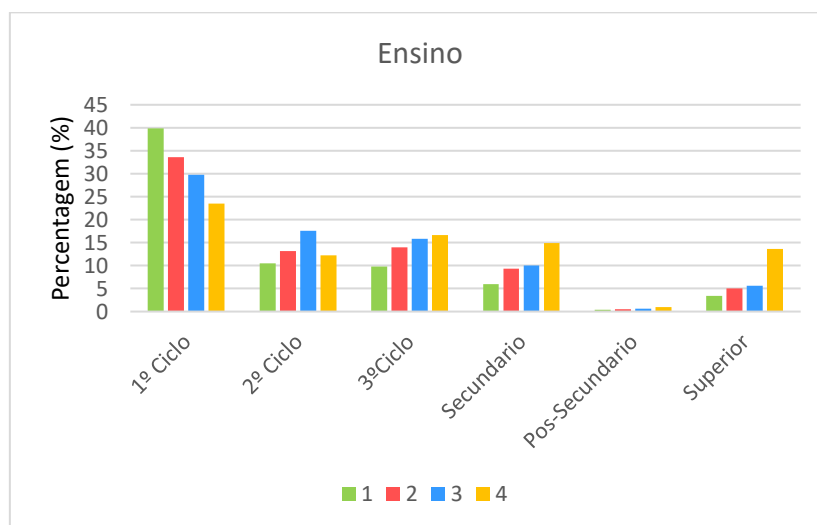


Gráfico 3.6 - Atributos referentes ao Ensino por cluster



## Empregabilidade

Em termos de empregabilidade, figura 3.9, as freguesias que apresentam um valor elevado de indivíduos residentes à procura de 1º emprego encontram-se maioritariamente nos distritos de Braga, Porto e Vila Real, sendo que é nestas que também existe uma forte presença de “High-High Clusters”. Ao considerar os indivíduos apenas desempregados à procura de emprego, as freguesias com percentagens elevadas deste atributo encontram-se mais numa zona litoral do Norte, com uma forte presença no distrito de Braga e do Porto, e nos distritos de Beja e Évora, e na zona litoral do distrito de Faro. As freguesias com elevada percentagem de indivíduos residentes empregados encontram-se numa zona litoral do Centro e Norte do continente, nomeadamente nos distritos de Aveiro, Braga, Coimbra, Leiria, Lisboa, Porto, Setúbal, sendo que nos distritos de Beja, Bragança, Castelo Branco, Guarda, Portalegre e Vila Real são os únicos distritos que não apresentam “High-High Clusters”. De certa forma podemos verificar que o distrito de Setúbal e Lisboa são uma zona que apresenta variedade neste grupo, uma vez que apresenta “High-High Clusters” seja tendo em conta indivíduos residentes desempregados ou empregados.

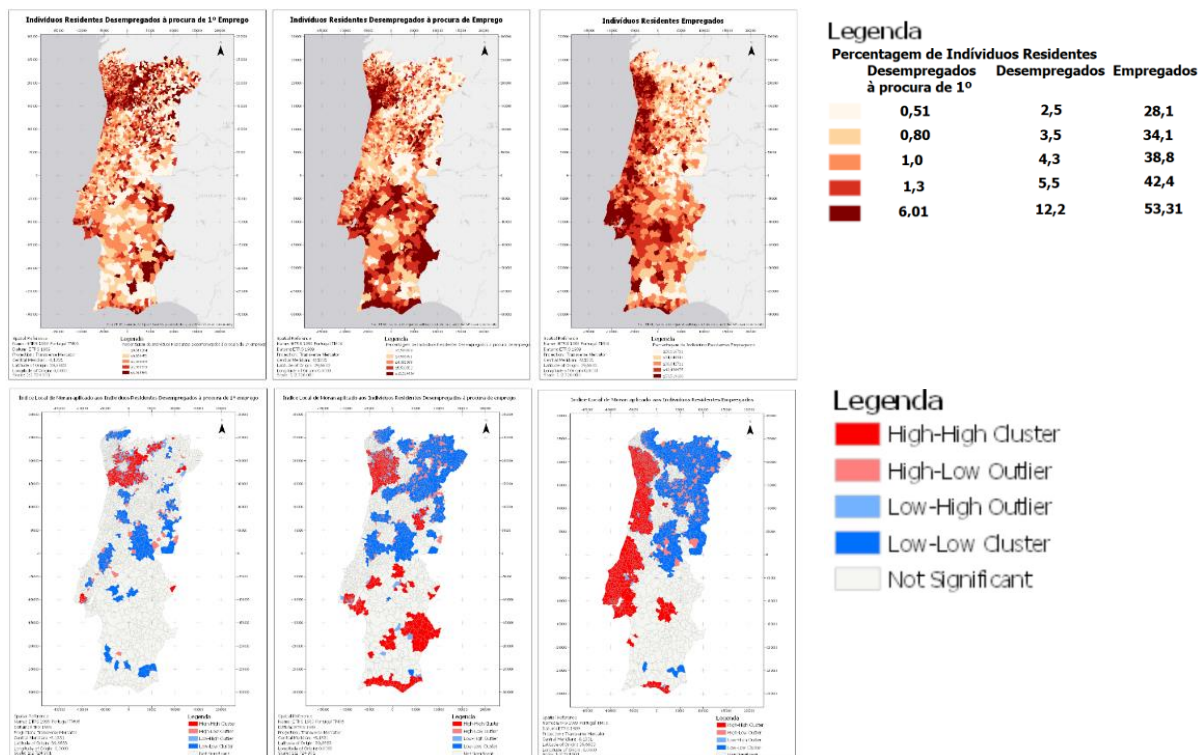


Figura 3.9 - Mapas dos atributos referentes à Empregabilidade ((da esquerda para a direita: Desempregados à procura de 1º emprego, Desempregados à procura de emprego e Empregados e respetivos índices locais de Moran

Em termos de indivíduos residentes Pensionistas e Reformados, ou Sem Atividade Económica, figura 3.10, através de ambos os mapas verificamos que o comportamento espacial é o oposto daquilo que foi apresentado na figura 3.9. A presença deste atributo é maior numa zona mais interior do país, existindo ainda algumas freguesias numa zona mais Sul do país, nomeadamente no distrito de Beja.

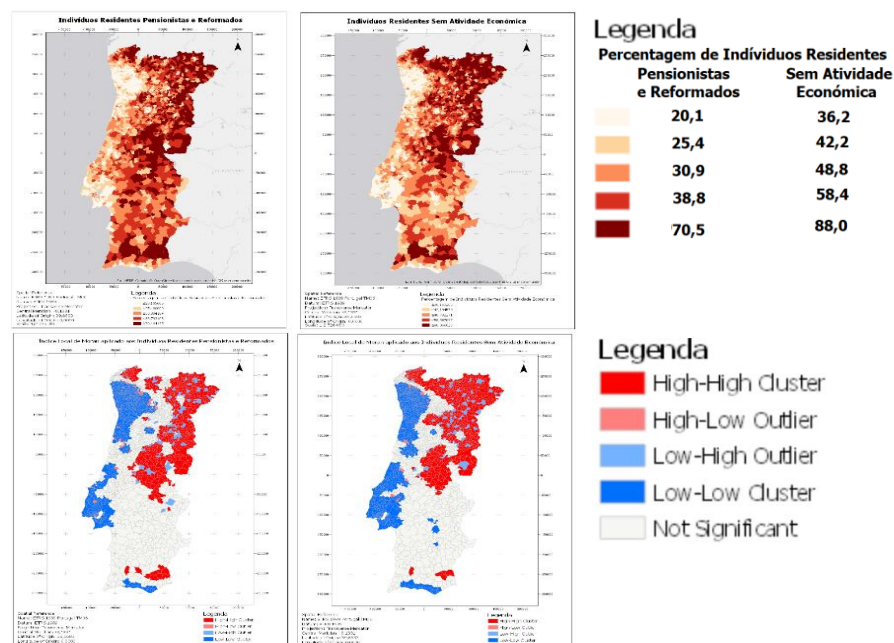


Figura 3.10 - Mapas dos atributos referentes à Empregabilidade (da esquerda para a direita: Pensionistas e Reformados e Sem Atividade Económica) e respetivos índices locais de Moran

Os indivíduos residentes Pensionistas e Reformados encontram-se predominantemente em maior percentagem no *cluster* 1, gráfico 3.7, o que confrontando o mapa com o índice de Moran destes dois atributos verificamos que os “High-High Cluster” encontram-se na mesma zona onde se encontra maioritariamente freguesias pertencentes a este *cluster*. Os indivíduos desempregados à procura de 1º emprego encontram-se em maior percentagem no *cluster* 3 e 4, apesar de serem diferenças mínimas percentuais. No entanto se analisarmos o mapa com o índice de Moran, verificamos que a zona de “High-High Cluster” corresponde a freguesias pertencentes a estes dois *clusters* apesar que também acaba por estar em freguesias pertencentes ao *cluster* 2. O mesmo acontece ao indivíduos desempregados à procura de emprego, apesar que uma vez que este tem um maior número de “High-High Cluster” na zona sul, nomeadamente no Alentejo e Algarve, em freguesias que pertencem ao *cluster* 2. Os indivíduos empregados pertencem em maior percentagem ao *cluster* 3 e 4, em que os “High-High Cluster” do mapa do índice de Moran se encontram nas freguesias maioritariamente pertencentes a estes *clusters*.

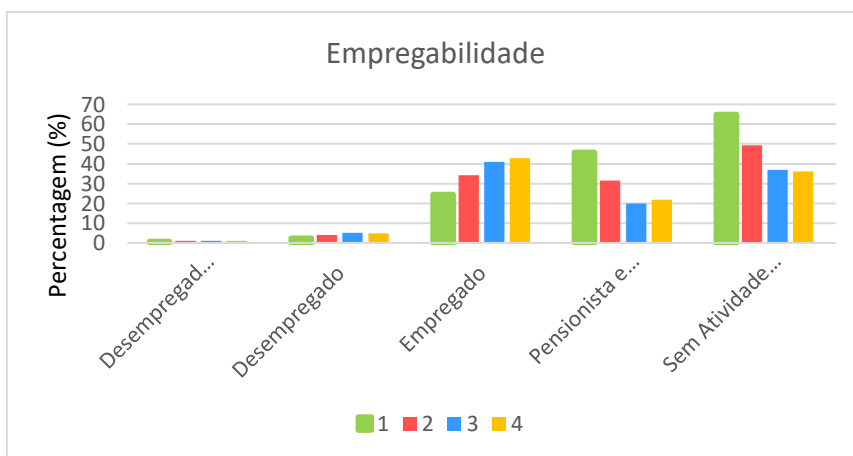


Gráfico 3.7 - Atributos referentes à Empregabilidade por cluster

### Rendimento

Sobre os atributos relacionados com o rendimento da população presente nas freguesias, as freguesias com valores percentuais altos de famílias com rendimentos pertencentes ao 1º quintil situam-se maioritariamente na zona interior Centro e Norte de Portugal o que mostra um comportamento espacial diferente de todos os restantes quintis, pois ao aproximarmo-nos ao litoral, a tendência é que as famílias pertençam a um quintil de rendimento superior, como se pode verificar na figura 3.11.

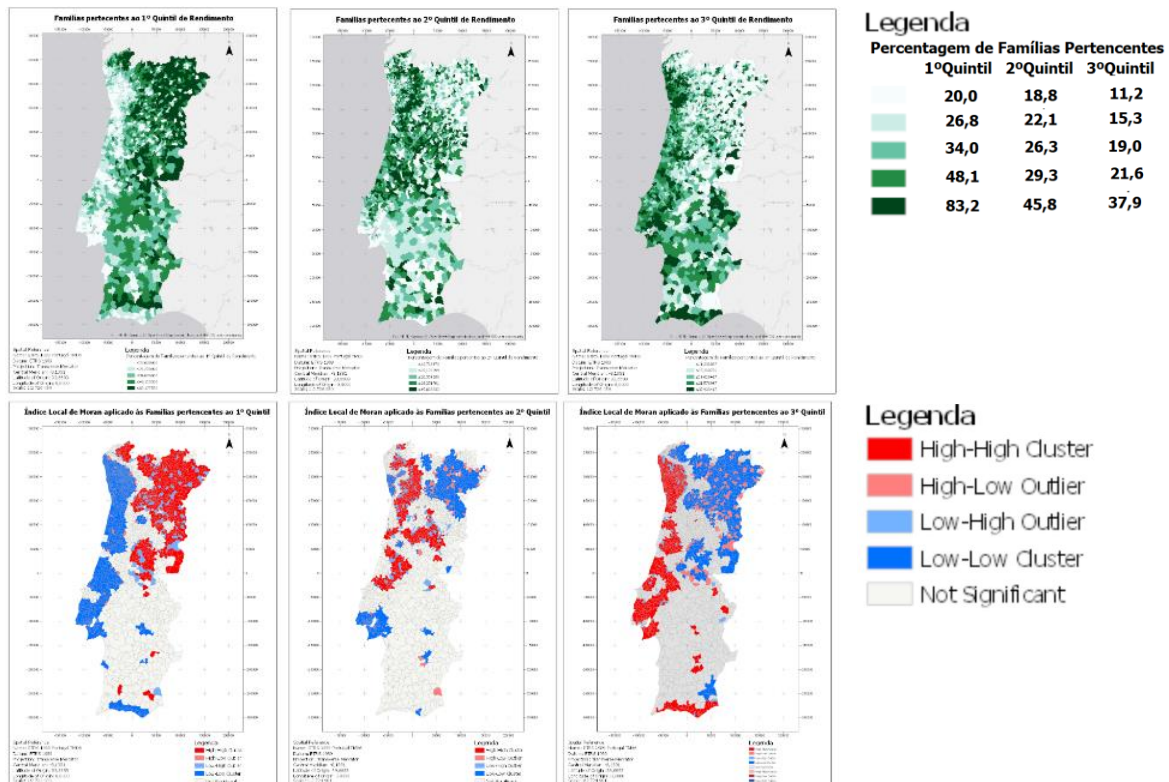


Figura 3.11 - Mapas dos atributos referentes ao Rendimento (da esquerda para a direita: 1º, 2º e 3º Quintil) e respetivos índices locais de Moran

Se considerarmos os dois últimos quintis de rendimento, ou seja, onde estão presentes os valores mais altos de rendimento, figura 3.12, as freguesias já se encontram distribuídas na zona litoral do continente, sendo que as freguesias onde a percentagem de famílias pertencentes ao 5º quintil de rendimento é alta encontram-se maioritariamente no distrito de Braga, Lisboa, Porto e Setúbal.

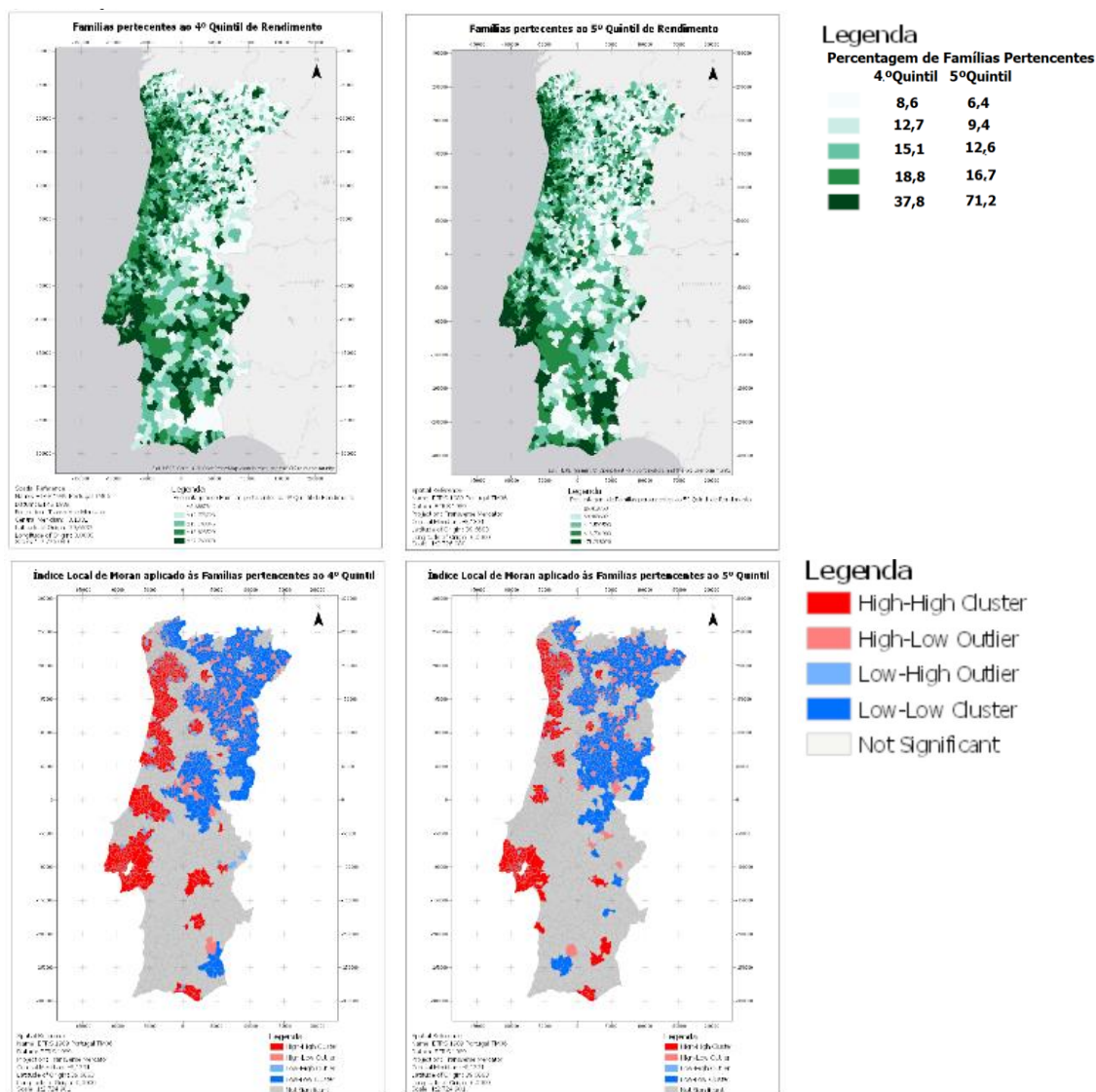


Figura 3.12 - Mapas dos atributos referentes ao Rendimento tendo em conta famílias pertencentes ao 4º e 5º Quintil e respetivos índices locais de Moran

Através do gráfico 3.8, podemos ver que no *cluster* 1 encontram-se maioritariamente as famílias pertencentes ao 1º quintil, e à medida que subimos nos quintis de rendimento, a percentagem de freguesias que pertencem ao *cluster* 1 vai diminuindo. Em contrapartida, o *cluster* 4 apresenta os maiores valores percentuais em níveis altos de rendimento.

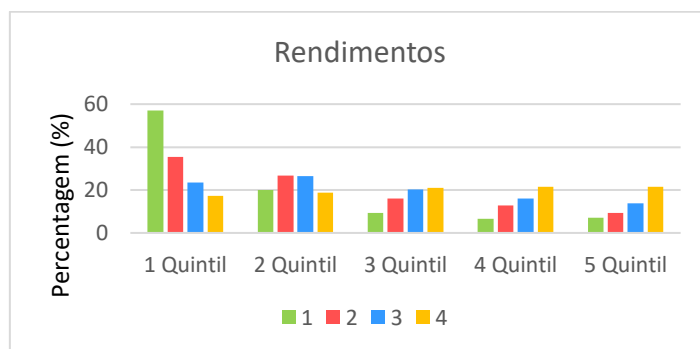


Gráfico 3.8 - Atributos referentes ao Rendimento por cluster



### Densidade Populacional

Na figura 3.13, as freguesias que têm valores altos de densidade populacional encontram-se situadas no distrito de Lisboa e do Porto, onde se encontram presentes os “High-High Cluster”, sendo que na zona interior do continente, de Norte a Sul, estão as freguesias com baixa densidade populacional, onde se encontra a maioria dos “Low-Low Cluster”.

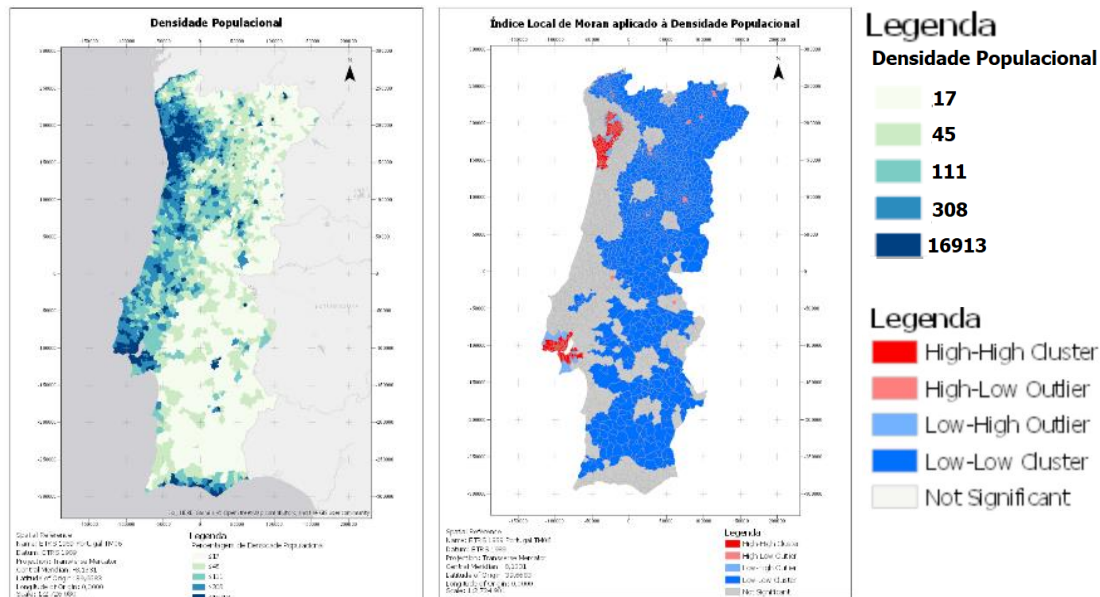


Figura 3.13 - Mapas da Densidade Populacional e respetivo índice local de Moran

Relativamente à atribuição das freguesias aos *clusters* segundo este atributo, gráfico 3.9, verificamos que as freguesias do *cluster* 4 são aquelas que apresentam elevados valores de densidade populacional, sendo que em seguida são as freguesias do *cluster* 3, depois do *cluster* 2. O *cluster* 1 é aquele que apresenta as freguesias com menor densidade populacional. O mapa com o índice de Moran corrobora estas afirmações uma vez que as zonas com “High-High Cluster” são maioritariamente freguesias pertencentes ao *cluster* 1, que é aquele que se destaca neste atributo por ter elevada densidade populacional.

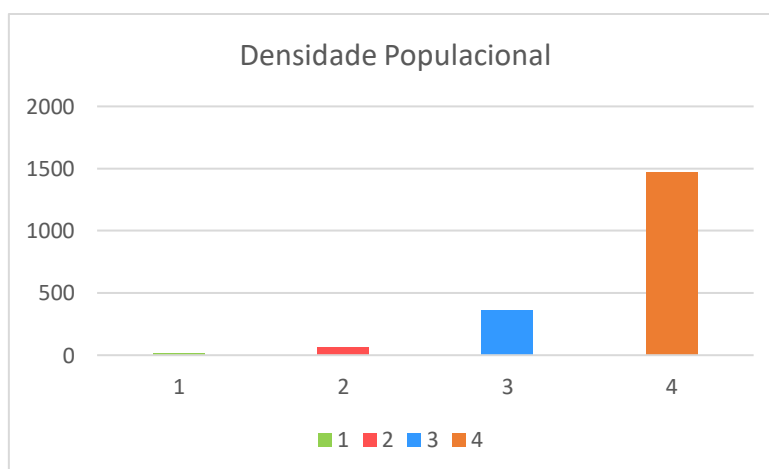


Gráfico 3.9 - Densidade Populacional por cluster

### Gasto em Alcool e Tabaco

Em termos de gasto em álcool e tabaco através da figura 3.14 verificamos que estes dois atributos se comportam de forma inversa em certas zonas do continente, mas em outras acaba por ser semelhante. Numa zona mais interior no Centro e Norte do continente é onde estão os maiores valores de gasto em Tabaco, sendo que numa zona mais Centro-Sul no litoral é onde existe maior gasto em álcool. No entanto nas regiões do Alentejo e Algarve existe a presença de “High-High Clusters” de ambos os atributos, embora exista um maior número de freguesias onde o consumo de Tabaco é alto nessas mesmas regiões.

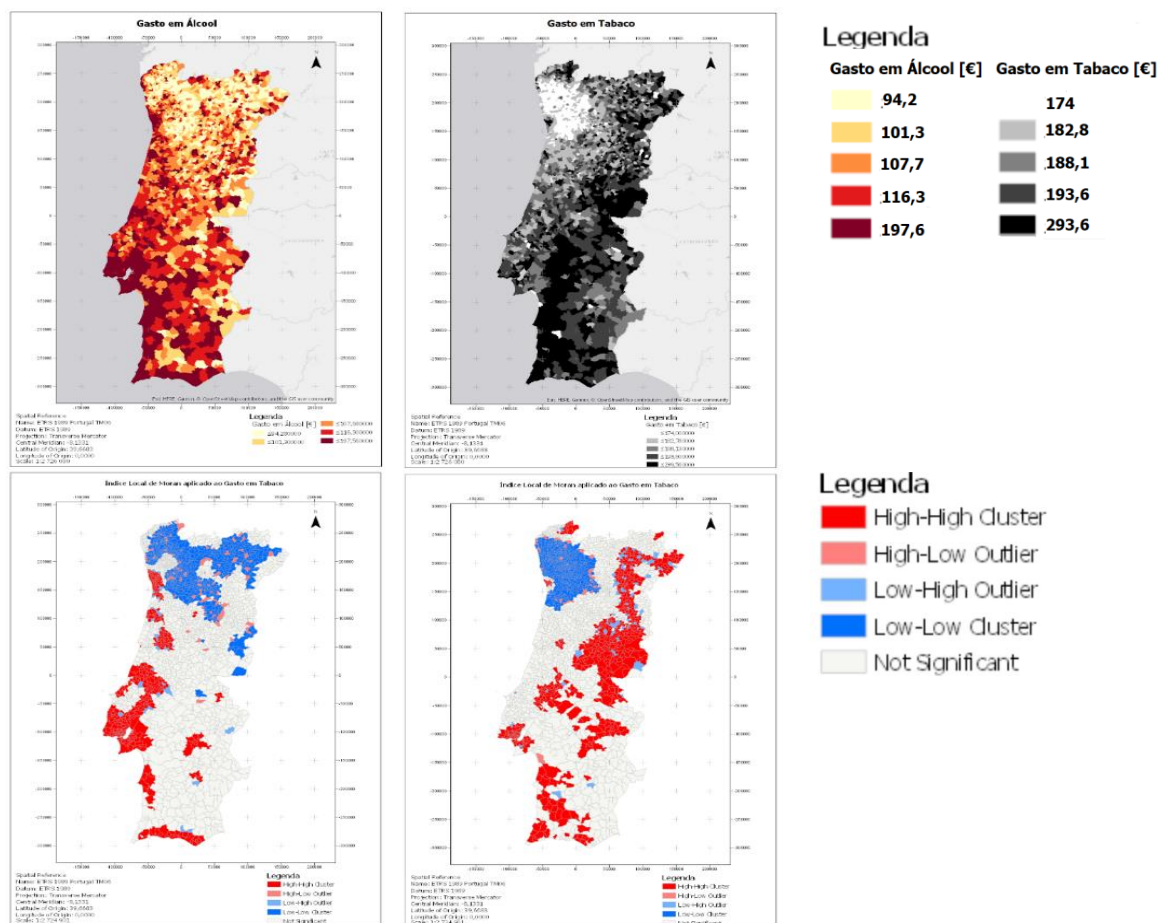


Figura 3.14 - Mapas do Gasto em Alcool e tabaco e respetivos índices locais de Moran

Em termos de gasto em álcool (gráfico 3.10) o *cluster* 4 é aquele que se destaca por serem freguesias onde se gasta mais 30€ que o *cluster* seguinte com valor mais alto, o *cluster* 2 com 103€, sendo que no *cluster* 3 existe um gasto de 102€ e no *cluster* 1 de 96,5€. Neste caso podemos verificar que as freguesias pertencentes ao *cluster* 1 encontram-se todas nas zonas de “High-High Cluster” enquanto que as de “Low-Low Cluster” são compostas por freguesias pertencentes aos restantes *clusters*.

Sobre o gasto em Tabaco (gráfico 3.10) já existe uma maior uniformidade nos valores entre o *cluster* 1, 2 e 4, apesar de, no entanto, o *cluster* 4 ser o que apresenta maior gasto, um total de 192€, sendo que o *cluster* 1 apresenta 191€, o *cluster* 2 de 186€ e o *cluster* 3 é aquele que apresenta menor valor, cerca de 171€. Se verificarmos o mapa do índice local de Moran, as zonas de “Low-Low Cluster” são as freguesias que pertencem ao *cluster* 3, reforçando a ideia transmitida pelo gráfico abaixo.

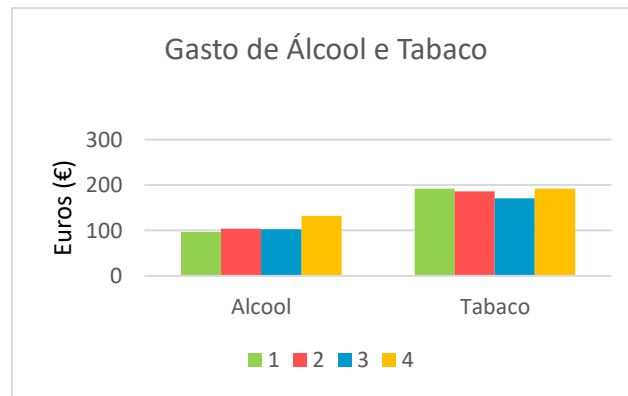


Gráfico 3.10 - Gasto em Álcool e Tabaco por cluster

### 3.3.2 Análise Espacial de Clusters e respetiva caracterização

Apesar desta segmentação ser feita com os mesmos atributos, a verdade é que cada *cluster* gerado terá dissemelhanças em relação aos outros *clusters* e algo que o caracteriza individualmente. Sendo que possa existir algum padrão espacial, é também provável que exista algum tipo de *outlier*, sendo que é nesse aspeto que o índice local de Moran é importante pois permite visualizar de forma rápida e eficiente qual é o atributo que faz com que aquela freguesia pertença a um *cluster* e não a outro. À semelhança do que foi realizado para os gráficos dos atributos espelhados pelos clusters, esta análise será feita através de comentários, e para cada *cluster* será feita a análise de um *outlier* de forma a tentar entender a razão da sua existência.

De referir que não será feita uma análise em termos de género dos indivíduos residentes pois em qualquer um dos *clusters* existe uma maior percentagem de indivíduos do género feminino do que masculino, como se verificou no gráfico 3.3 (página 35).

#### Cluster 1

Em termos espaciais, na figura 3.15, podemos verificar que as 603 freguesias pertencentes a este *cluster* se encontram maioritariamente numa zona interior do Centro-Norte, como por exemplo, no distrito de Bragança, Castelo Branco, Guarda e Vila Real, apesar de existirem também numa zona mais sul, no distrito de Beja e Faro.

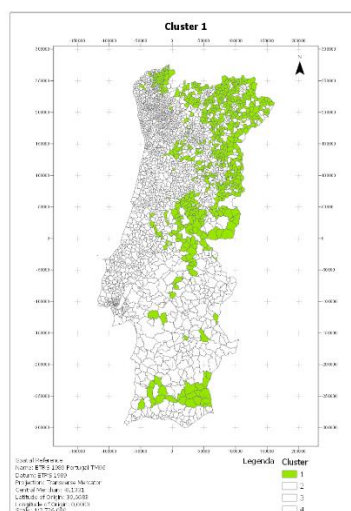


Figura 3.15 - Mapa com a localização das freguesias pertencentes ao Cluster 1

Através do gráfico 3.11 verificamos que a faixa etária com maior percentagem é a faixa etária dos 25 aos 64 anos, com 43%, seguida da faixa etária dos 65 ou mais anos com 42%. Tendo em conta o gráfico 3.5 (página 37), verificamos que este *cluster* é aquele que apresenta uma maior percentagem dos indivíduos com a faixa etária dos 65 ou mais anos.

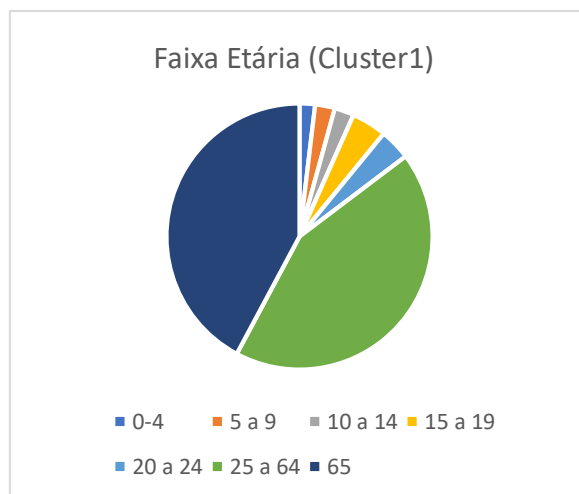


Gráfico 3.11 - Faixa Etária (Cluster 1)

Os indivíduos residentes pensionistas e reformados são aqueles que apresentam uma maior percentagem neste *cluster*, cerca de 47% gráfico 3.12, seguido por indivíduos sem atividade económica, sendo que a percentagem de desempregados neste caso é mínima, um total de 2%.

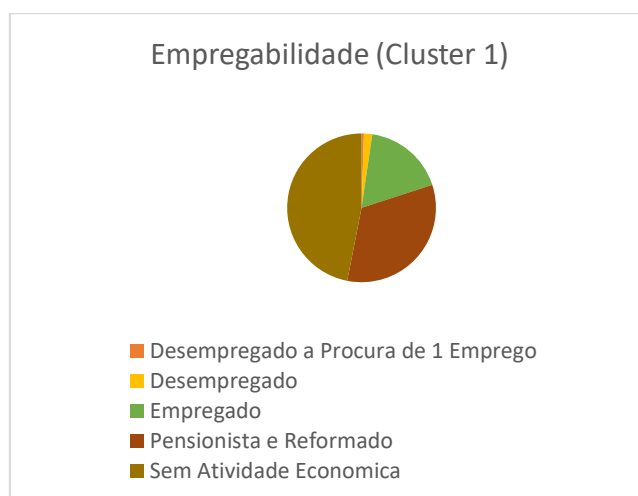


Gráfico 3.12 - Empregabilidade (Cluster 1)

No gráfico 3.13, a maior percentagem de indivíduos residentes em termos de ensino encontra-se naqueles que têm o 1º ciclo completo, sendo que o restante se encontra dividido de forma semelhante, estando 15% nos que completaram o 2º ciclo, 14% nos que completaram o 3º ciclo, 9% o ensino secundário, 5% completaram o ensino superior e cerca de 1% o ensino pós-secundário.

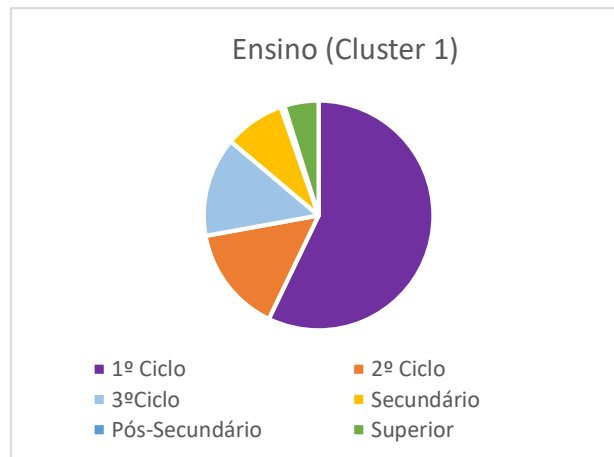


Gráfico 3.13 - Ensino (Cluster 1)

As famílias presentes nas freguesias deste *cluster* encontram-se maioritariamente no 1º quintil de rendimento, cerca de 57% (gráfico 3.14). Sendo que 20% das freguesias fazem parte do 2º quintil de rendimento, encontrando-se o resto dividido pelos restantes 3 quintis de rendimento.

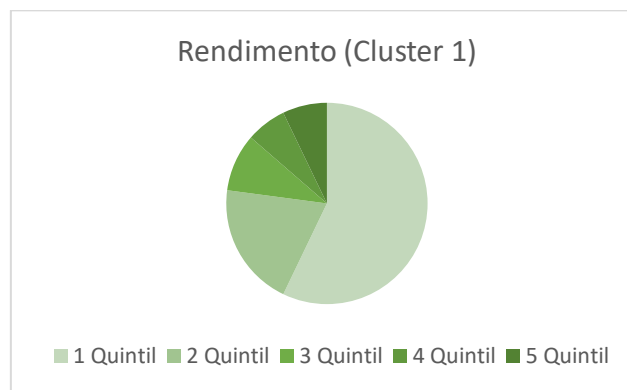


Gráfico 3.14 - Rendimento (Cluster 1)

A partir da análise anterior feita individualmente a cada atributo e tendo em conta o que foi analisado nesta secção, o *cluster* 1 é onde se encontra a maioria das pessoas idosas, que são pensionistas e reformadas e que têm baixos rendimentos, podendo-se afirmar que são pessoas que acabam por pertencer a uma classe baixa. Na maioria são pessoas que têm um baixo nível de escolaridade, tendo na maioria o 1º e o 2º ciclo completo. São áreas em que por norma existe um alto gasto em Tabaco em vez de Álcool comparativamente às restantes zonas. Estes indivíduos por norma vivem no interior do continente, onde existe pouca população, e numa zona mais a Norte do país, sendo que, no entanto, ainda existe alguma presença no Alentejo.

Como exemplo de *outlier* deste *cluster* foi escolhido a freguesia “União de freguesias de Monte e Queimadela” no distrito de Braga. Para entender quais os atributos que possam justificar a atribuição desta freguesia a este *cluster* é necessário recorrer ao índice local de Moran e verificar em quais destes atributos esta freguesia é considerada um *outlier*. Existem dois tipos de *outlier*, os “High-Low Outlier”, com uma cor vermelha suave, em que a presença deste significa que nessa mesma área o atributo tem um valor alto, mas nas áreas circunjacentes o valor do atributo é baixo, sendo que neste caso deverão ser os atributos que melhor caracterizam este *cluster* e o difere dos outros. Os “Low-High Outlier”, com uma cor azul suave, são áreas em que o atributo terá um valor reduzido, sendo que nas áreas circunjacentes o valor será elevado, o que significa que esse mesmo atributo caracteriza os *clusters* que

se encontram junto a este *outlier*. De notar que neste caso estamos a fazer uma análise local, o que significa que apesar de uma freguesia estar alocada a um *cluster* com certas características, não significa que não tenha algumas características de outros *clusters*. O resultado do algoritmo de *clustering* é a combinação de todas as variáveis e não apenas de uma.

Na figura 3.16 está a localização deste *outlier* e um exemplo de ambos os tipos de *outlier*.

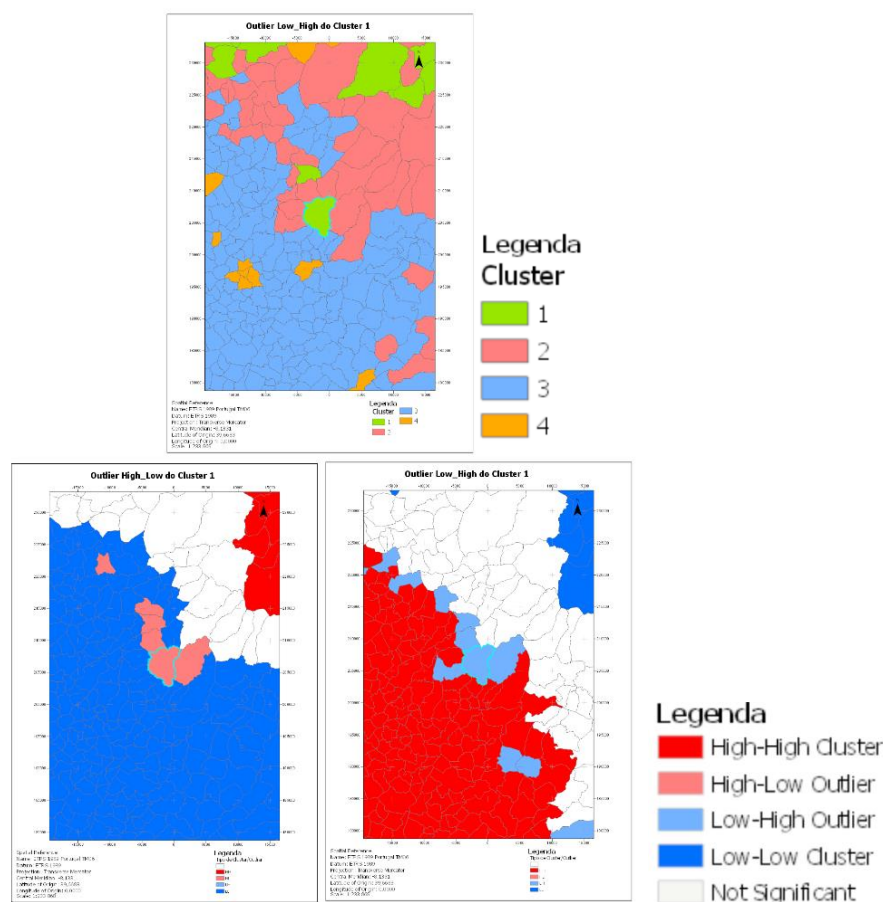


Figura 3.16 - Mapa com exemplo de Outlier do Cluster 1 e a análise com índice local de Moran do atributo “65 ou mais anos de idade” e “Faixa Etária 25 a 64 anos”

Verificando todos os mapas com o índice local de Moran para todos os atributos, nesta freguesia, os atributos que fazem com que esta freguesia seja considerada um “High-Low Outlier” são: Faixa Etária de 65 ou mais anos de idade (exemplo da figura), Indivíduos Residentes Sem Atividade Económica e Pensionistas e Reformados e Famílias presentes no 1º quintil de Rendimento. O exemplo que se encontra na figura 3.16 comprova que o atributo “Faixa Etária de 65 ou mais anos” é algo que caracteriza e distingue esta freguesia de, por exemplo, as freguesias que estão perto do *cluster* 3.

Os atributos que fazem com que a freguesia seja considerada um “Low-High Outlier” são: Faixa etária de 5 a 9 anos, 10 a 13 anos, 14 a 19 anos, 20 a 24 anos, 25 a 65 anos (exemplo da figura) e Indivíduos Residentes Desempregados à procura de emprego. Neste caso verifica-se que nesta freguesia não existem indivíduos pertencentes à faixa etária abaixo dos 65 anos, caracterizando esta freguesia com uma população mais idosa.

Cluster 2

Apesar de este ser o *cluster* que tem um maior número de freguesias, com um total de 1052, é aquele que apresenta um padrão espacial menos uniforme, figura 3.17. A maioria das freguesias do distrito de Beja, Évora, Santarém pertencem a este *cluster*.

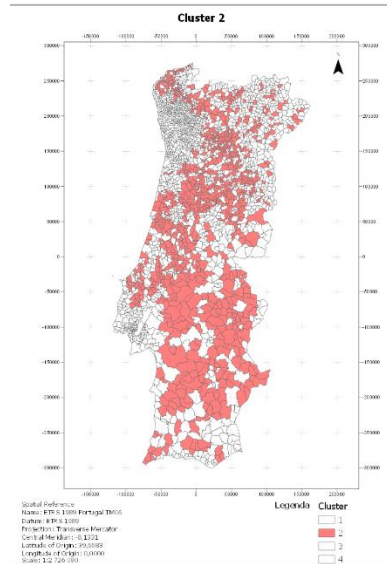


Figura 3.17 - Mapa com a localização das freguesias pertencentes ao Cluster 2

Comparativamente ao *cluster* anterior, podemos verificar que neste, apesar de ainda existir uma grande percentagem de indivíduos com 65 ou mais anos, cerca de 28%, gráfico 3.15, a maior percentagem corresponde aos indivíduos que tem entre 25 a 64 anos, sendo que as restantes faixas etárias correspondem a no máximo 6% dos indivíduos.

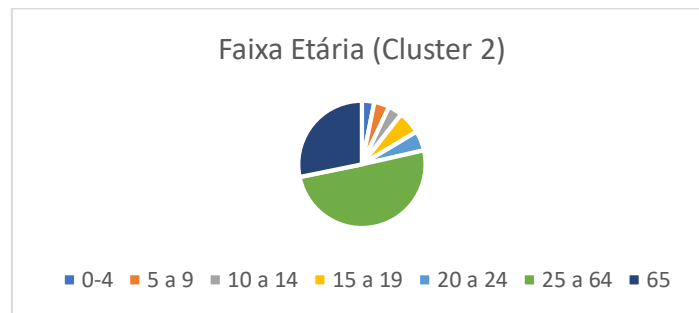
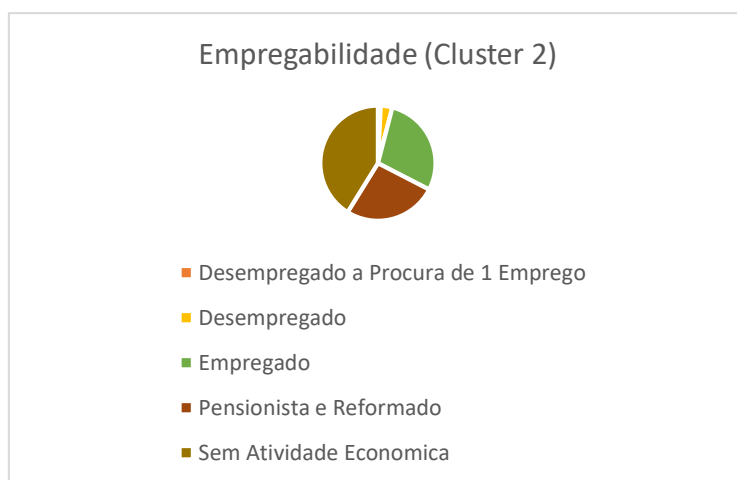


Gráfico 3.15 - Faixa etária (Cluster 2)

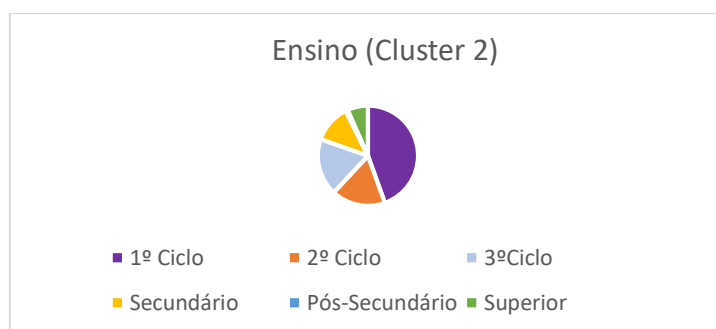
Relativamente a empregabilidade, gráfico 3.16, neste *cluster* continuam a predominar os indivíduos que não tem atividade económica, no entanto existe uma grande percentagem de empregados, cerca de 28%, e de pensionistas e reformados, cerca de 26%, enquanto que a percentagem de indivíduos desempregados é cerca de 4% no total.





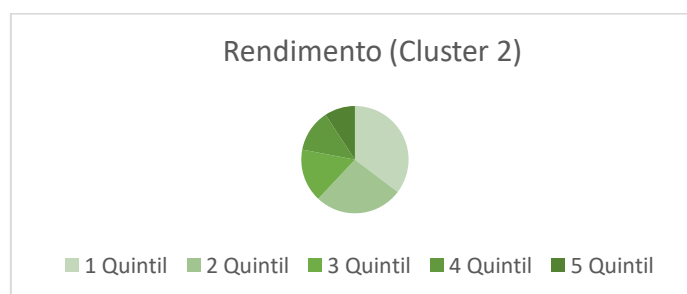
*Gráfico 3.16 - Empregabilidade (Cluster 2)*

Nestas freguesias, verificando o gráfico 3.17, predominam os indivíduos com o 1º ciclo completo, cerca de 45%, e que tenham o 2º ciclo, 17%, e o 3º ciclo completo, 18%. Cerca de 12% dos indivíduos completaram o secundário e 7% completaram o ensino superior, sendo que o ensino pós-secundário não atinge 1% dos indivíduos presentes na freguesia.



*Gráfico 3.17 - Ensino (Cluster 2)*

O 1º e o 2º Quintis de rendimento continuam a predominar neste *cluster*, gráfico 3.18, correspondendo conjuntamente a 62% das famílias presentes nesta freguesia.



*Gráfico 3.18 - Rendimento (Cluster 2)*

De certa forma, em termos de atributos, este *cluster* tem algumas semelhanças com o *cluster* 1, a população residente nesta freguesias acaba por ser uma população mais adulta e madura, com uma forte presença nas faixas etárias entre 25 e 65 e 65 ou mais anos. São pessoas que, por norma, não tem um elevado grau de escolaridade, no entanto já conseguem ter algum rendimento, apesar de não elevado, podendo de certa forma estar inseridos numa classe média-baixa, comparativamente aos restantes



*clusters*. São freguesias que não são grandes centros urbanos, pois o valor de densidade populacional não é de todo elevado.

Neste caso, a título de exemplo, o *outlier* escolhido foi a freguesia “Vinhais” no distrito de Bragança, que se apresenta na figura 3.18.

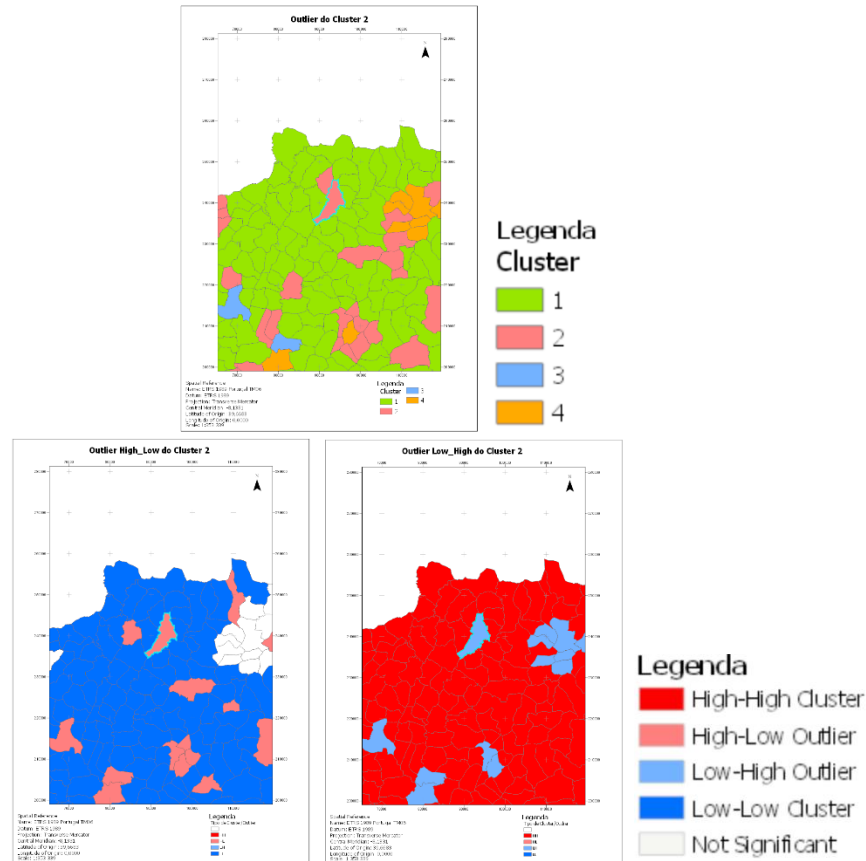


Figura 3.18 - Mapa com exemplo de Outlier do Cluster 2 e a análise com índice local de Moran

Neste caso, os atributos que fizeram com que esta freguesia fosse considerada um “High-Low Outlier” foram: Género Feminino, Faixa Etária de 25 a 64 anos (exemplo da figura), Ensino do 3º Ciclo, Secundário e Ensino Superior Completo e famílias pertencentes ao 4º Quintil de Rendimento. Como afirmado, este *cluster* é semelhante ao *cluster* 1, cor verde na imagem, só que apresenta uma maior percentagem de população com uma faixa etária mais baixa que os 65 anos de idade e com um melhor nível de rendimento e escolaridade ligeiramente melhor.

Considerando a freguesia como um “Low-High Outlier”, os atributos que influenciaram este resultado foram: Género Masculino, Faixa Etária de 65 ou mais anos, Ensino do 1º Ciclo completo, Indivíduos Residentes Sem Atividade Económica e Pensionistas e Reformados e famílias pertencentes ao 1º Quintil de Rendimento. Todos estes atributos foram utilizados para caracterizar o *cluster* 1.

### Cluster 3

Existe uma grande concentração de freguesias pertencentes a este *cluster*, com um total de 755 freguesias, no Norte, nos distritos de Braga, Porto e Vila Real, figura 3.19, apesar de existirem algumas freguesias com características semelhantes em outras zonas do país, mas não de uma forma tão aglomerada.

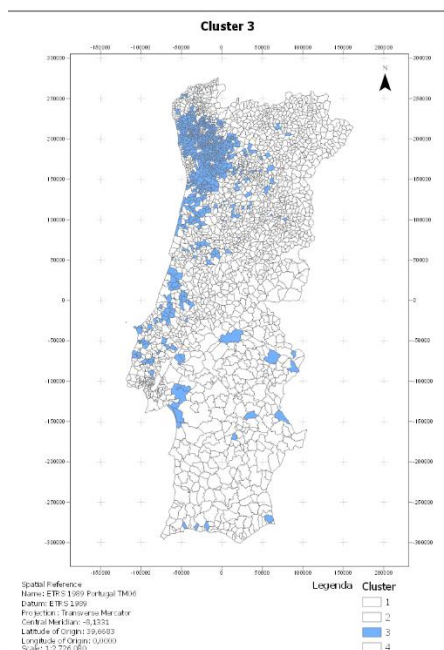


Figura 3.19 - Mapa com a localização das freguesias pertencentes ao Cluster 3

A grande maioria dos indivíduos deste *cluster* encontra-se numa faixa etária de 25 a 64 anos, cerca de 55% (gráfico 3.19). No entanto é de relembrar que na análise no subcapítulo anterior (página 34), é neste *cluster* que se encontram as maiores percentagens dos indivíduos mais novos, com faixa etária de 0 a 14 anos, como se verificou no gráfico 3.4 (página 37).

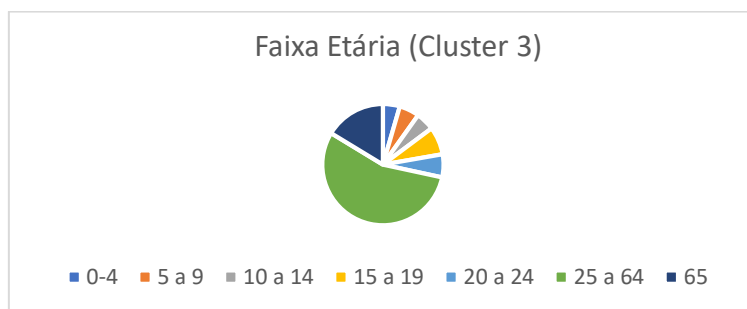


Gráfico 3.19 - Faixa Etária (Cluster 3)

Neste *cluster*, gráfico 3.20, já existe uma maioria nos indivíduos empregados, correspondendo a cerca de 39%, sendo seguido de 36% de indivíduos sem atividade económica. No entanto, e como afirmado anteriormente, ao analisar o gráfico 3.7 (página 41), este *cluster* é aquele que apresenta uma maior percentagem de indivíduos desempregados, correspondendo a 5% dos indivíduos residentes nas freguesias.

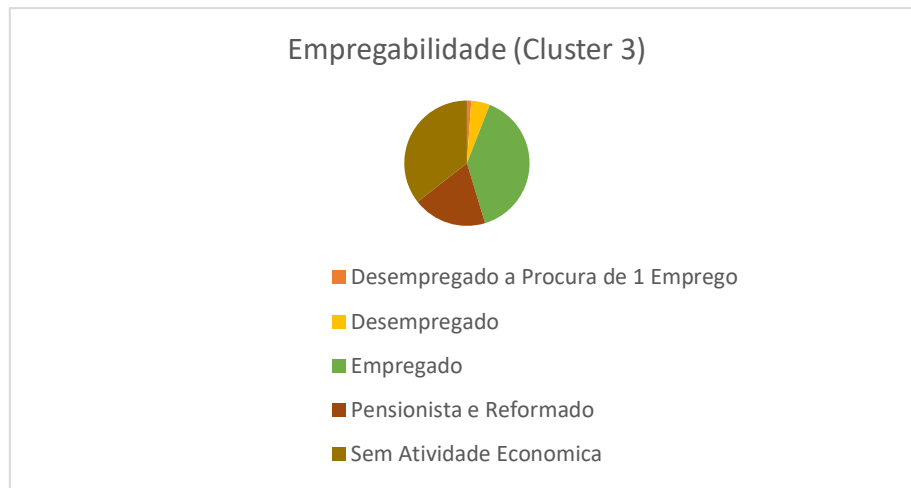


Gráfico 3.20 - Empregabilidade (Cluster 3)

Apesar de ainda existir uma predominância de indivíduos com apenas o 1º ciclo completo, a sua percentagem já é menor comparativamente aos *clusters* anteriores, sendo de 38%, como se pode verificar no gráfico 3.21. As percentagens de indivíduos residentes com o 2º e 3º ciclo completo já são maiores, o que em parte faz sentido uma vez que é neste *cluster* que existe maior percentagem de crianças e adolescentes.

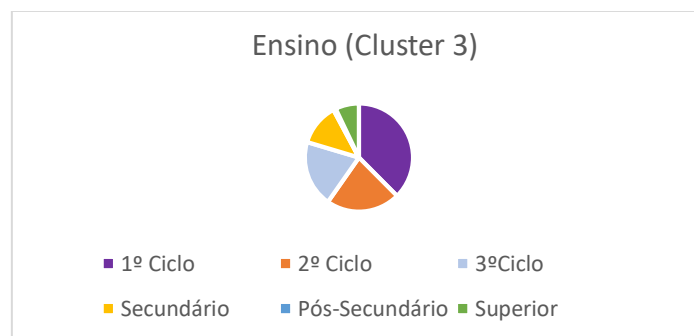


Gráfico 3.21 - Ensino (Cluster 3)

Em termos de rendimento, gráfico 3.22, este *cluster* já apresenta menores percentagens de famílias pertencentes ao 1º e 2º quintis de rendimento comparativamente aos *clusters* anteriores, apresentando já valores mais elevados nos restantes quintis, demonstrando que neste *cluster* já existe um nível de rendimento maior que nos anteriores. Verifica-se que continuam a predominar quintis de níveis mais baixo de rendimento, nomeadamente o 1º, 2º e 3º, sendo 79% das famílias residentes.

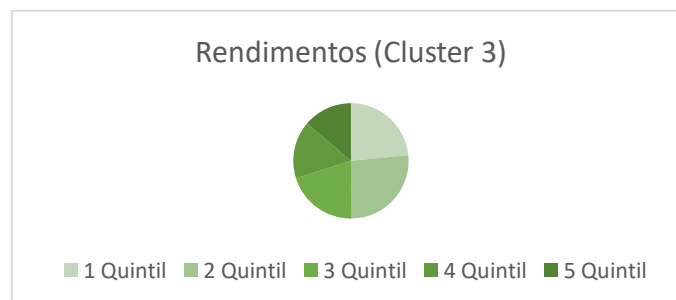


Gráfico 3.22 - Rendimentos (Cluster 3)

Neste *cluster* estão presentes os indivíduos residentes com uma faixa etária mais baixa, sendo que é onde se encontram aqueles que provavelmente ainda irão terminar o ensino secundário e que estão a ingressar o ensino superior. Acabam por ser uma população mais jovem, no entanto as famílias já apresentam algum nível de rendimento, podendo ser considerados numa classe média comparativamente aos restantes *clusters*.

A freguesia “Carvalhais” no distrito de Bragança foi utilizada como exemplo de *outlier* deste *cluster*, figura 3.20.

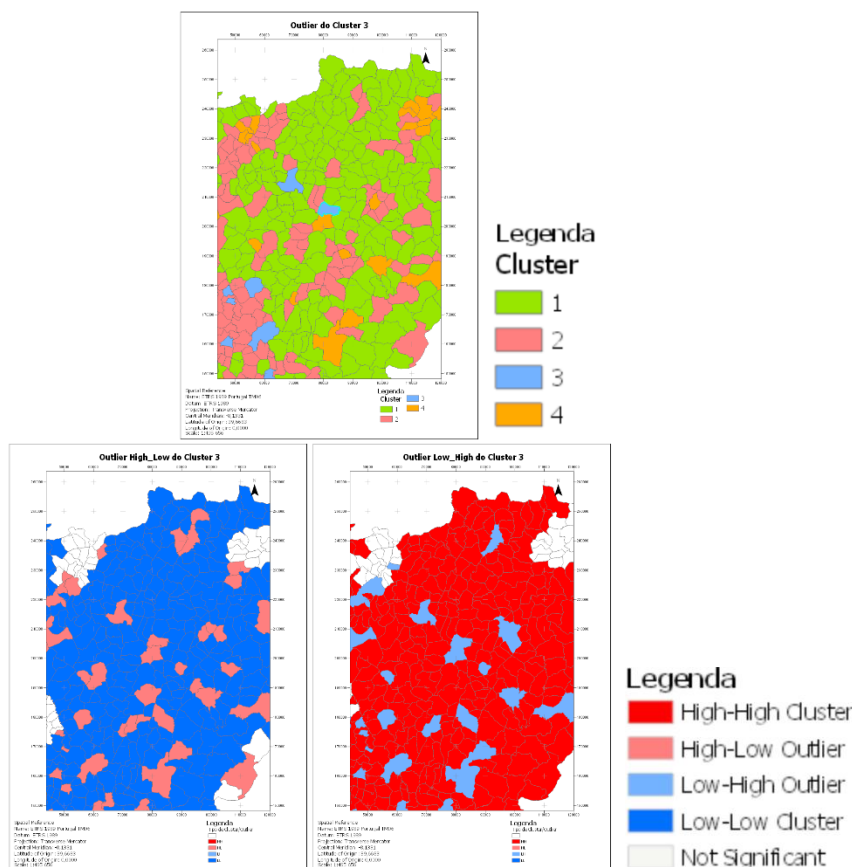


Figura 3.20 - Mapa com exemplo de Outlier do Cluster 3 e a análise com índice local de Moran

Neste caso, os atributos que tiveram influência para que esta freguesia fosse considerada um “High-Low Outlier” foram: Faixa Etária de 5 a 9 anos, 20 a 24 anos, 25 a 64 anos, Ensino de 3º Ciclo, Secundário e Superior completo e Famílias pertencentes ao 3º Quintil de Rendimento (exemplo da figura 3.20). Podemos verificar que os atributos utilizados para caracterizar este *cluster* encontram-se aqui nesta localização, um nível de escolaridade a nível do ensino secundário e superior, uma faixa etária de jovens e jovens adultos e algum nível de rendimento em comparação a, por exemplo, às freguesias do *cluster* 1 e 2, que é onde se encontram indivíduos com uma faixa etária mais elevada e um nível de rendimentos mais baixo, por norma no 1º e 2º Quintil de Rendimento.

Se considerarmos esta freguesia como um “Low-High Outlier”, os atributos influenciadores foram: Faixa Etária de 65 ou mais anos, Ensino do 1º Ciclo completo, Indivíduos Residentes Sem Atividade Económica e Pensionistas e Reformados e Famílias pertencentes ao 1º Quintil de Rendimento. Como anteriormente foi dito, estas são características do *cluster* 1, que é o tipo de *cluster* que se encontra mais presente nas freguesias que rodeiam este *outlier*.

Cluster 4

As freguesias presentes neste *cluster*, com um total de 472 freguesias, figura 3.21, são aquelas que se encontram em grandes centros urbanos, seja nomeadamente Lisboa, Coimbra, Porto, entre outros. Existe uma grande concentração de freguesias pertencentes a este *cluster* no distrito de Lisboa e Setúbal e na zona litoral do distrito de Faro, no entanto é possível verificar algumas freguesias espalhadas ao longo do continente.

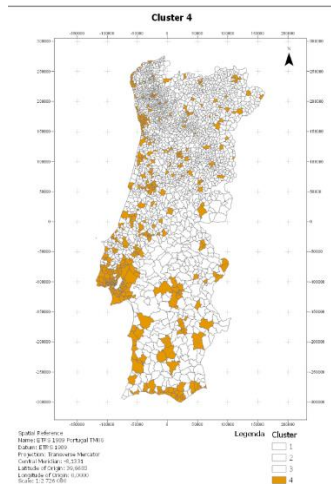


Figura 3.21 - Mapa com a localização das freguesias pertencentes ao Cluster 4

A faixa etária dos 25 aos 65 anos predomina nas freguesias deste *cluster*, gráfico 3.23, sendo que em seguida encontram-se os indivíduos com 65 ou mais anos.

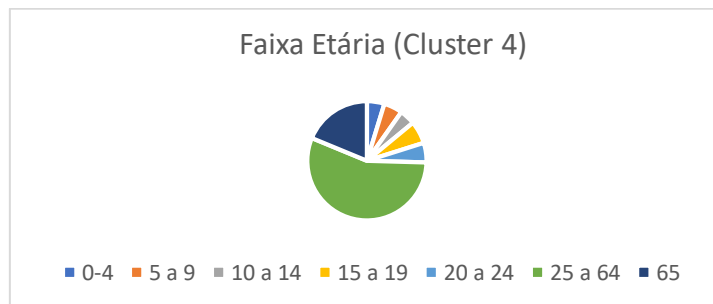


Gráfico 3.23 - Faixa Etária (Cluster 4)

Neste caso, no gráfico 3.24, os indivíduos empregados predominam nas freguesias, com cerca de 40%, sendo seguidos de indivíduos sem atividade económica com 34%.

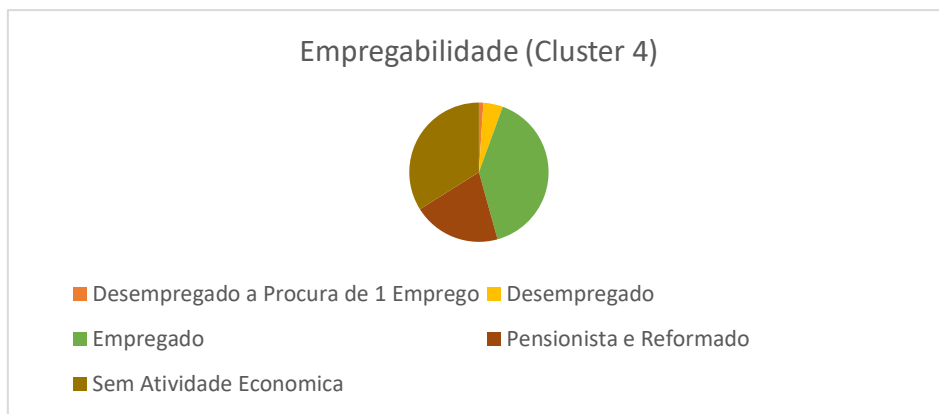


Gráfico 3.24 - Empregabilidade (Cluster 4)

Apesar de os indivíduos com o 1º ciclo completo serem os que se encontram em maior percentagem no gráfico 3.25, seguidos dos que têm o 3º ciclo completo, os indivíduos com o ensino secundário e superior completo já apresentam uma percentagem maior comparativamente com os *clusters* anteriores, nomeadamente de 18% e 17%

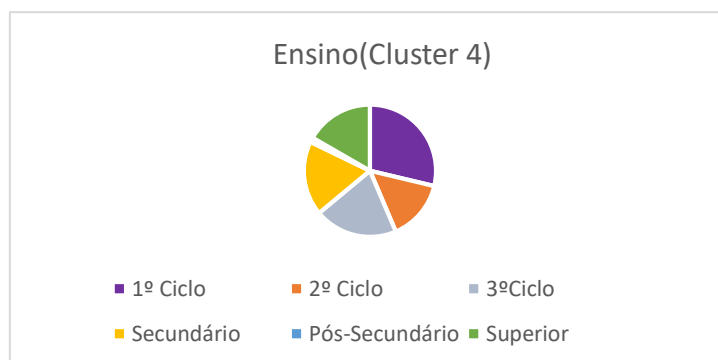


Gráfico 3.25 - Ensino (Cluster 4)

As famílias pertencentes ao 4 e 5º quintis de rendimento encontram-se em maioria no gráfico 3.26, com uma percentagem 22% ambas, totalizando quase a maioria, apesar de os restantes quintis se encontrarem com valores não muito dispares.

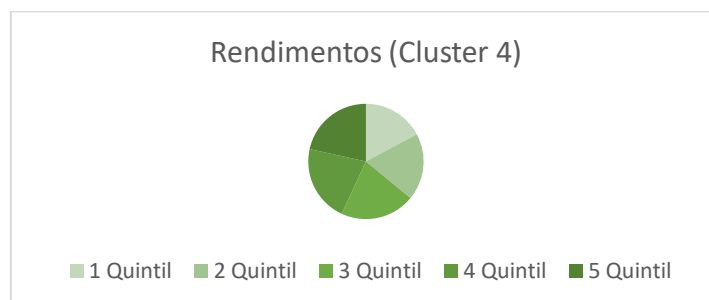


Gráfico 3.26 - Rendimentos (Cluster 4)

Os indivíduos presentes neste *cluster* são jovens trabalhadores, com um grau de escolaridade elevado, como por exemplo o ensino superior completo, sendo que grande parte destes se encontram empregados e residem em grandes centros urbanos, pois a densidade populacional é elevada. É nestas zonas onde existe um maior gasto em álcool e tabaco.

A freguesia selecionada como exemplo de *outlier* para este *cluster* foi a freguesia “Castelo Branco” no distrito de Castelo Branco, figura 3.22.

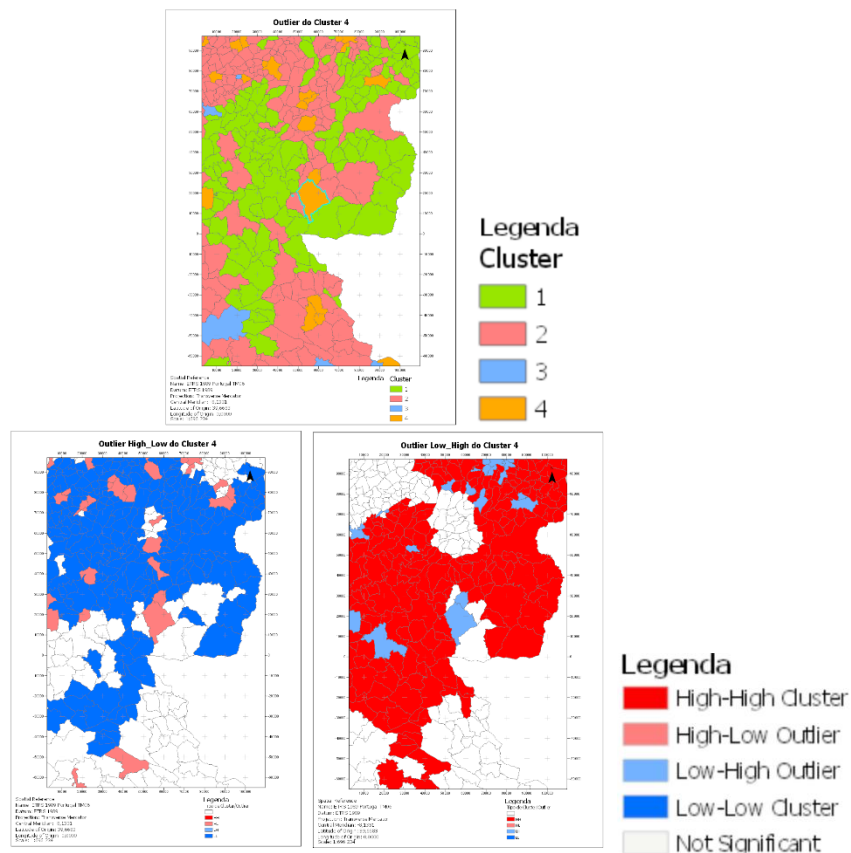


Figura 3.22 - Mapa com exemplo de Outlier do Cluster 4 e a análise com índice local de Moran

Neste caso, para a freguesia ser considerado “High-Low Outlier” os atributos a ter em conta foram: Faixa Etária de 10 a 13 anos, 14 a 19 anos e 20 a 24 anos, Indivíduos Residentes Desempregados à procura de emprego e Famílias pertencentes ao 5º Quintil de Rendimento (exemplo da figura 3.22). Estes atributos estão de acordo com as características analisadas e escolhidas para este *cluster* numa análise anterior.

No caso de a freguesia ter sido considerada um “Low-High Outlier”, os atributos que influenciaram foram: Faixa Etária de 65 ou mais anos (exemplo da figura 3.22), Indivíduos Residentes Desempregados à procura de emprego. Neste caso como o atributo utilizado como exemplo é uma característica do *cluster* 1 e algumas das freguesias vizinhas são pertencentes a este mesmo *cluster*, e como referido anteriormente, no *cluster* 4 existem maiores percentagens de indivíduos jovens adultos é normal que esta freguesia apresente este tipo de outliers no atributo “Faixa Etária de 65 ou mais anos”.

### 3.4 Análise Exploratória de dados de diabetes

Nesta fase do projeto procurou-se tirar partido dos resultados da segmentação e de uma análise de autocorrelação espacial e efetuar uma análise espacial conjunta com os dados de diabetes. Os dados de diabetes são os dados provenientes da base de dados da Transparência [43], no conjunto de dados relativos a “Atividade do Programa de Diabetes”. Neste conjunto encontra-se o número de utentes inscritos com diabetes com exame de pés realizado no último ano e o número de utentes inscritos com diabetes com último resultado de HgbA1c inferior ou igual a 8% e as respetivas proporções desses atributos, ou seja, apresenta o número de utentes diabéticos inscritos que demonstram estar preocupados

com a sua saúde ao efetuarem exames de controlo sobre a sua doença. Estes dados têm uma série temporal de quatro anos, respetivamente 2014, 2015, 2016 e 2017.

Uma vez que este conjunto de dados se encontrava a um nível geográfico de ACES, foi necessário criar um ficheiro de polígonos que contivesse a localização de cada uma destas áreas. Como referido anteriormente, existem cerca de 55 ACES a nível continental em Portugal. A partir da informação disponibilizada pelo SNS foi construída uma camada de informação de polígonos com a totalidade dos ACES [44]. A maioria dos ACES são responsáveis por um ou mais municípios, sendo que alguns na zona de Lisboa e Porto são responsáveis apenas por um conjunto de freguesias, nomeadamente, Grande Porto V, Grande Porto VI, Grande Porto VII, Grande Porto VIII, Lisboa Central, Lisboa Norte e Lisboa Ocidental e Oeiras. Esta camada de informação foi criada tendo como base o ficheiro da CAOP 2016. Tendo em conta informação disponibilizada pelo SNS, foi selecionado o município, o conjunto destes ou freguesias, se for o caso, e foi aplicada a ferramenta de geoprocessamento “*Dissolve*” efetuado 55 vezes, uma para cada ACES. Após a conclusão deste processo obteve-se a camada de informação de polígonos que está representada na figura 3.23.

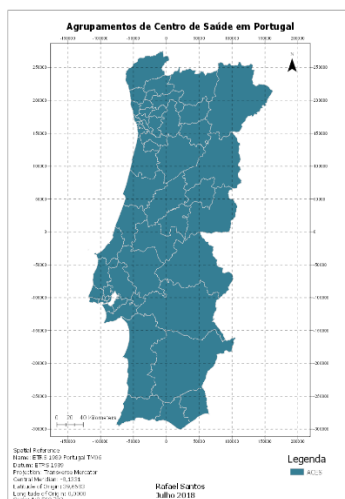


Figura 3.23 – Mapa com os Agrupamentos de Centros de Saúde (ACES) em Portugal

Uma vez que estes dados têm uma componente temporal foi necessário criar uma camada de informação para cada ano. Primeiramente recorreu-se à ferramenta de geoprocessamento *Select Layer by Attribute*, sendo que a layer escolhida foi a que continha os dados dos diabetes, e a expressão utilizada foi “tempo contains the text 2014”, no caso de 2014. Foi feito um *Export Table* para criar uma nova camada de informação com dados apenas referente a esse ano, por exemplo, a camada de informação denominada “Diabetes\_2014”, sendo esta uma camada de informação de pontos com informação referente a cada ACES em cada mês desse ano com dados de diabetes. Uma vez que queremos a informação organizada por ACES de forma anual, foi necessário utilizar o *Summarize* na tabela de atributos de forma a poder diminuir o nível de informação temporal por ACES. A *Input Table* era a tabela “Diabetes\_2014” e em todos os atributos foi utilizada a *Statistic Type* “Sum” sendo que o *Case Field* é o atributo “ACES”, o atributo com o nome do ACES. O último passo foi efetuar um *Join* entre esta tabela e a camada de informação “ACES”, representada na figura 3.23, para juntar esta informação aos polígonos que representam cada ACES e recorrer ao *Export Features* para criar uma nova camada denominada, por exemplo, por “ACES\_Diabetes\_2014”. Este processo foi efetuado para os quatro anos presentes nos dados.



Como referido anteriormente, os dados sobre a diabetes presentes neste projeto são uma série temporal, de 2014 a 2017, do número de utentes inscritos com diabetes com exame de pés realizado no último ano e o número de utentes inscritos com diabetes com último resultado de HgbA1c inferior ou igual a 8% e as respetivas proporções desses atributos, ou seja, qual a proporção de pessoas que têm diabetes mas que demonstram estar preocupados com a sua saúde ao efetuarem exames de controlo sobre a sua doença. A principal análise será feita tendo em conta os valores de proporções, pois o número de utentes inscritos apenas nos dá uma noção de onde existem mais utentes inscritos que realizam este tipo de exames, não demonstra onde é que existem mais utentes que sejam diabéticos e demonstrem preocupação ao terem a sua doença controlada, pois as proporções são relativas a utentes que sejam diabéticos mas que tenham exames realizados ou com valores saudáveis em comparação ao total de utentes diabéticos. A distribuição espacial destes dados encontram-se presentes no Anexo II.

O ACES Baixo Mondego e ACES Baixo Vouga são aqueles que ao longo dos 4 anos apresentaram o maior número de utentes inscritos que realizaram o exame do pé, como se pode observar através da figura 3.24.

É no próximo passo que ferramentas como a segmentação ou o Índice Local de Moran são importantes para tentar entender se existe algum atributo que justifique este comportamento. Uma vez que estamos a falar em número de utentes, como a segmentação está feita com características internas da freguesia, e não é feita uma comparação de atributos a nível nacional, não é possível verificar, por exemplo, onde existem mais indivíduos do género masculino a nível nacional, mas sim quais as freguesias que têm maiores percentagens de indivíduos do género masculino. Neste caso o único atributo que poderia justificar a existência de um elevado número de utentes seria a Densidade Populacional, e, no entanto, não seria o mais adequado pois este demonstra o número de indivíduos residentes por área. Contudo, através da densidade populacional verificamos que por norma as zonas mais populadas se encontram numa zona litoral do continente, justificando que a zona interior do país se encontre em tons de vermelho ao longo dos anos nos mapas, pois significa que é onde existe um número menor de utentes inscritos com este exame realizado.

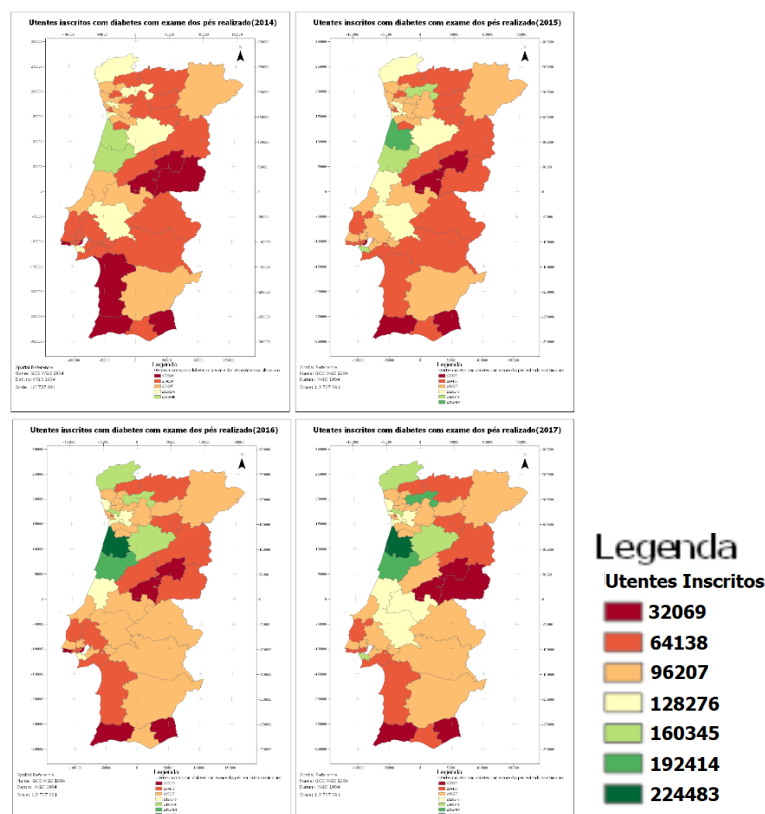


Figura 3.24 - Série temporal de Utentes inscritos com diabetes com exame dos pés realizado (2014, 2015, 2016 e 2017)

Ao analisarmos os valores de proporção deste mesmo atributo já podemos ter em consideração características da população nas áreas geográficas, pois estes valores são referentes a população diabética que controla a sua doença comparativamente ao total de população diabética, e este tipo de comportamento poderá, ou não, estar associado a características da população. Através da figura 3.25, verificamos que as áreas em que existem maior número de utentes não significa que seja onde exista uma maior percentagem. Exemplo disso é o ACES Baixo Vouga que em todos os anos foi o ACES que apresentou maior número de utentes inscritos a realizar o exame, no entanto não é aquele que apresenta os maiores valores de proporção. Na zona do litoral Norte, nomeadamente no distrito de Porto (ACES Grande Porto II, ACES Grande Porto III, ACES Grande Porto IV, ACES Grande Porto VIII) e Aveiro (ACES Entre Douro e Vouga I - Feira e Arouca) é onde existe desde 2014 os níveis de proporção mais altos, entre 57 a 72% dos utentes inscritos diabéticos, sendo que existe sempre uma evolução ao longo dos anos nesta zona, juntando-se mais alguns do distrito do Porto (ACES Grande Porto V, ACES Matosinhos), do distrito de Aveiro (ACES Entre Douro e Vouga II), do distrito de Braga (ACES Cávado I, ACES Cávado III, ACES Ave – Famalicão, ACES Alto Ave - Guimarães, Vizela e Terras de Basto) e do distrito de Vila Real (ACES Douro I). Se verificarmos o mapa com o resultado da segmentação, figura 3.3 (página 33), grande parte das freguesias desta área são freguesias alocada ao *cluster* 3, o que significa que a população nesta área é uma população mais jovem mas com algum nível de rendimento e escolaridade, ou seja, significa que têm algumas possibilidades financeiras de se poderem deslocar e efetuarem os exames e capacidade de terem acesso à informação, uma vez que hoje em dia e em faixas etárias mais baixas já existe uma mentalidade e destreza diferente de encarar as plataformas, principalmente as digitais, que fornecem a informação.

Numa zona interior do Centro-Norte, dois ACES, um no distrito da Guarda (ACES Guarda) e outro no distrito de Castelo Branco (ACES Cova da Beira) não alteram o seu comportamento ao longo destes 4 anos, mantendo uma percentagem de utentes diabéticos inscritos entre os 14% e os 28%. Se

tivermos em conta os dados da figura 3.24 (figura acima), o número de utentes inscritos no ACES de Castelo Branco e da Guarda tem-se mantido constante, sendo que o da Guarda é dos ACES que apresenta um menor número de utentes inscritos que realizaram este exame. De certa forma isto demonstra que uma vez que este número se tem mantido constante, e os valores de proporção também, não ocorreu uma mudança de comportamento nos utentes inscritos, sendo que provavelmente os que realizaram os exames, continuaram a realizá-los nos anos seguintes e os que não realizaram, não o efetuaram nos restantes anos. Nestes dois ACES encontram-se maioritariamente freguesias pertencentes ao *cluster* 1 e 2, tendo estes uma população mais envelhecida, com menos rendimentos e um nível de escolaridade baixo, o que pode significar que não tenham recursos para efetuar exames, seja em termos de deslocação ou informação para compreenderem a importância do rastreio, ou como são na maioria zonas com baixa densidade populacional, o centro de saúde mais próximo não seja necessariamente na sua freguesia. De forma geral, a proporção de utentes inscritos diabéticos a realizar este exame tem vindo a melhorar ao longo dos anos, sendo que partir de 2016, o mínimo considerado seria entre 14% a 28%, valores mais elevados que nos anos anteriores, sendo apenas 5 o número de ACES que se mantêm com esses valores de proporção em 2017.

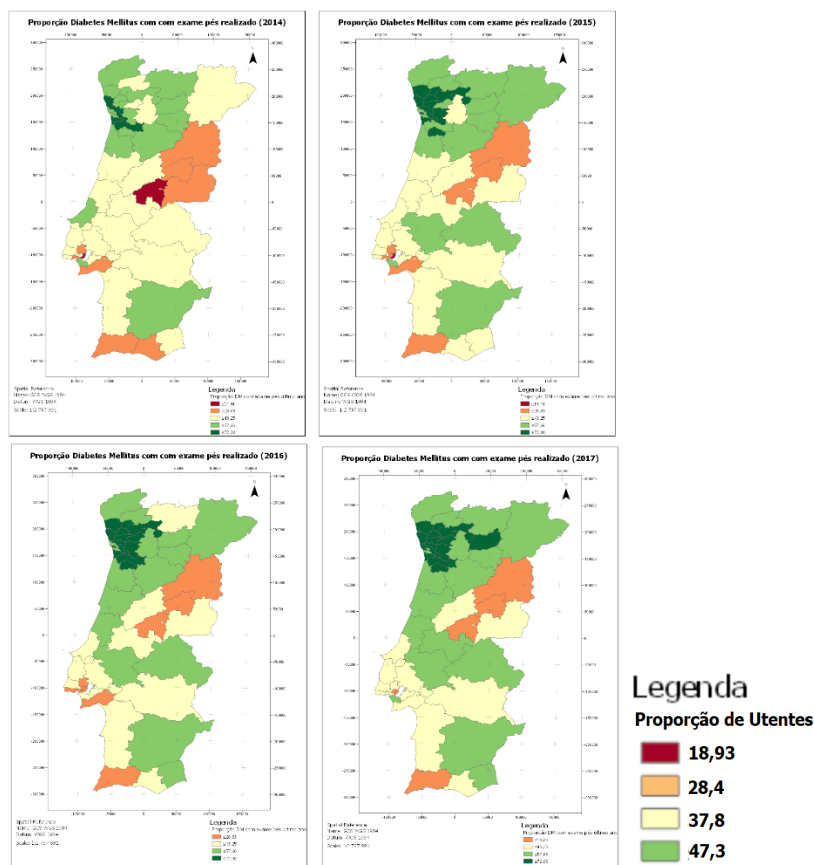


Figura 3.25 - Série temporal de Proporção de Utentes inscritos com diabetes com exame dos pés realizado (2014, 2015, 2016 e 2017)

Também podemos ver que apenas dois ACES apresentam uma percentagem abaixo dos 14% em 2014, cor vermelho escuro, sendo o ACES Pinhal Interior Sul e o ACES Lisboa Central. Se verificarmos as freguesias que se encontram na área do ACES Lisboa Central, todas estas pertencem ao *cluster* 4. Então o que poderá justificar o fato deste ACES apresentar valores mais reduzidos na proporção? Através dos mapas gerados com o índice local de Moran podemos verificar quais os atributos que diferenciam alguma das freguesias. Na figura 3.26 encontra-se representada a localização deste ACES e quais os atributos que influenciam a que exista “High-Low Outliers” nesta área geográfica, pois são estes que

nos vão dizer as freguesias em que o valor do atributo é elevado, mas que na sua vizinhança os valores são reduzidos. Após analisar todos os resultados do índice Local de Moran em todos os atributos, verificou-se que aqueles que demonstravam a presença dos *outliers* referidos anteriormente foram a “Faixa Etária de 65 ou mais anos” e “Pensionistas e Reformados”. De certa forma podemos ver que isto está interligado com o que anteriormente foi analisado, os piores valores de proporção estão em zonas onde a presença de freguesias do *Cluster 1* é maior, e ambos os atributos são características deste mesmo *cluster*.

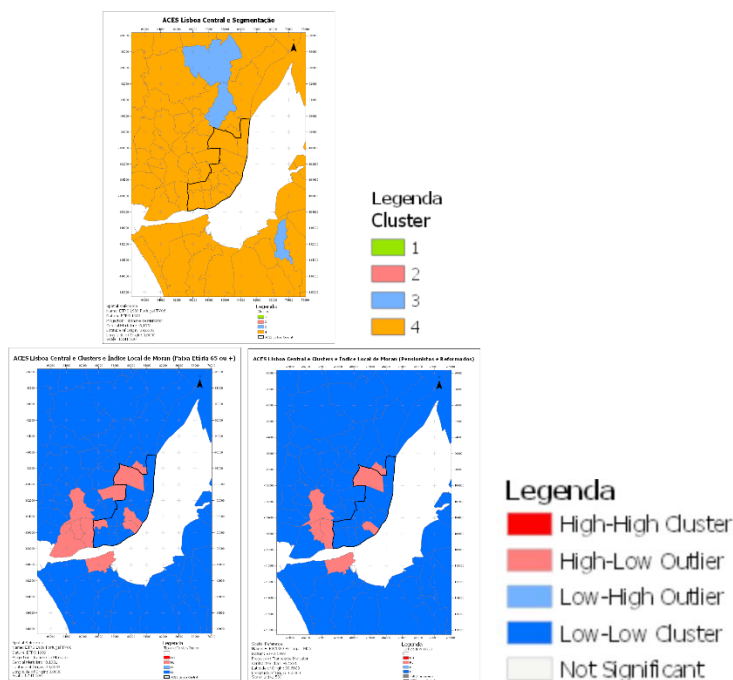


Figura 3.26 - ACES Lisboa Central com resultados da segmentação e índices locais de Moran a alguns atributos

Podemos verificar através da figura 3.27 que o padrão espacial dos utentes diabéticos inscritos com o último resultado de HgbA1c igual ou inferior a 8% é semelhante ao padrão daqueles que realizaram o exame do pé. Em parte, poderá ser porque ambos exames deverão ter que ser realizados, continuando a verificar-se que os ACES do Baixo Vouga do distrito de Aveiro a apresentar um maior número de indivíduos, sendo que já se verifica um número inferior nos distritos do Porto. No distrito de Castelo Branco é onde existe um padrão constante de um número reduzido de utentes nesta zona. Em contrapartida, na zona litoral Sul do continente nos primeiros anos existia um número reduzido, mas depois acabou por aumentar no ano de 2017.

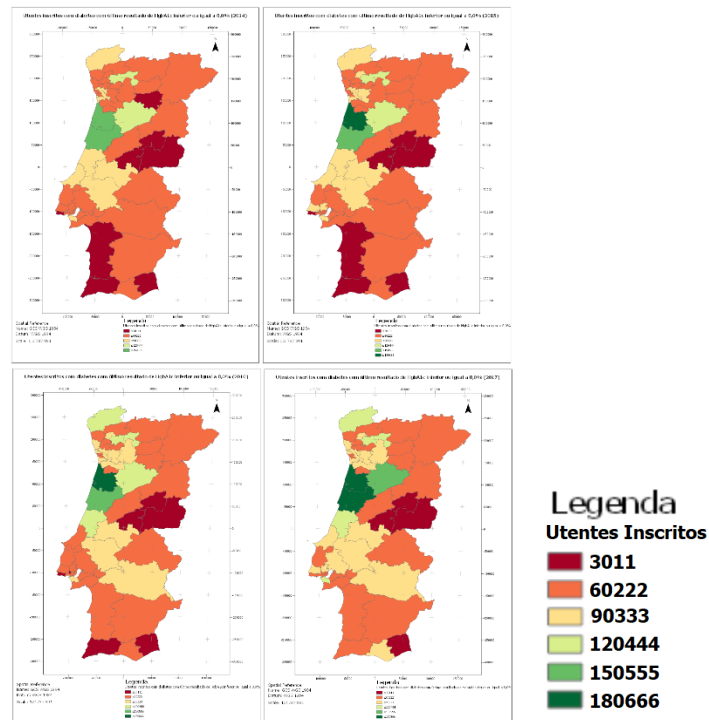


Figura 3.27 - Série temporal de Utentes inscritos com diabetes com último resultado de HgbA1c inferior ou igual a 8,0% (2014, 2015, 2016 e 2017)

Tendo em conta estes utentes inscritos com o resultado do exame de HgbA1c igual ou inferior a 8%, mas em proporção, figura 3.28, verificarmos que a zona litoral Norte, nomeadamente os ACES do distrito de Aveiro, Braga, Porto, Vila Real e Viseu, é a zona onde existe maior percentagem de utentes com este resultado no exame, por norma entre os 37 e 47% dos utentes têm este resultado. No entanto, o ACES Cávado I do distrito de Braga é aquele que desde 2015 apresenta a maior percentagem deste atributo, com valores entre os 47% e 56%. De forma geral, os resultados dos utentes inscritos têm melhorado, uma vez que no ano de 2014 existiam ACES que tinham uma percentagem inferior a 18%, mas no ano de 2017 já existia apenas ACES com um mínimo de percentagem entre os 18% e 28%.

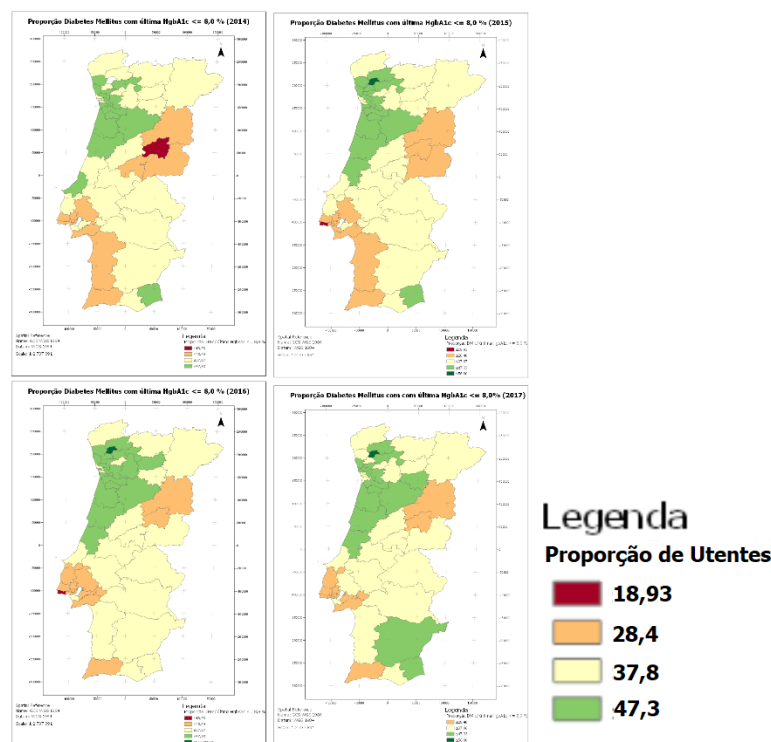


Figura 3.28 - Série temporal de Proporção de Utentes inscritos com diabetes com último resultado de HgA1c inferior ou igual a 8,0% (2014, 2015, 2016 e 2017)

De forma a entender o porquê de o ACES Cavado I se destacar a nível nacional, vamos recorrer aos resultados da segmentação e também aos resultados do índice local de Moran nos atributos. Desta forma vamos procurar “High-Low Outliers” e “Low-High Outliers” presentes na área do ACES ou na sua vizinhança que possa justificar o porquê de este apresentar um comportamento diferente dos ACES vizinhos. Através da figura 3.29, podemos verificar que neste ACES estão presentes freguesias pertencentes ao *cluster* 3 e ao *cluster* 4. Ambos têm como característica uma população jovem adulta, com bons rendimentos e nível de escolaridade. Neste ACES é onde existe uma concentração de freguesias do *cluster* 4, e este é o *cluster* onde existe maior percentagem de indivíduos com o ensino superior completo e com maior número famílias pertencentes ao 4 e 5º quintil de rendimento, o que poderá significar de certa forma que são indivíduos com possibilidades de se deslocar a entidades de saúde e com capacidade de ter acesso a informação. Quanto ao índice local de Moran, é possível verificar que os atributos “5º Quintil de Rendimento”, “4º Quintil de Rendimento”, “Ensino Superior” e “Ensino Secundário” são aqueles que apresentam alguns “Low-High Outliers” na vizinhança deste mesmo ACES, o que significa que o ACES Cavado I tem um valor mais elevado destes atributos do que algumas freguesias da sua vizinhança, pois por norma dentro da sua área encontram-se “High-High Cluster”. Quanto ao atributo “1ºCiclo”, no mapa é possível ver que algumas freguesias dos ACES vizinhos apresentam “High-Low Outliers”, o que significa que nestas freguesias existe um valor elevado deste atributo, enquanto que no ACES Cavado I encontram-se apenas “Low-Low Cluster” o que significa que nesta zona o atributo tem valores reduzidos. Desta forma, é possível entender quais os atributos que contribuíram para que este ACES se destaque dos outros, através de atributos que influenciam os ACES vizinhos e não propriamente o ACES em causa.



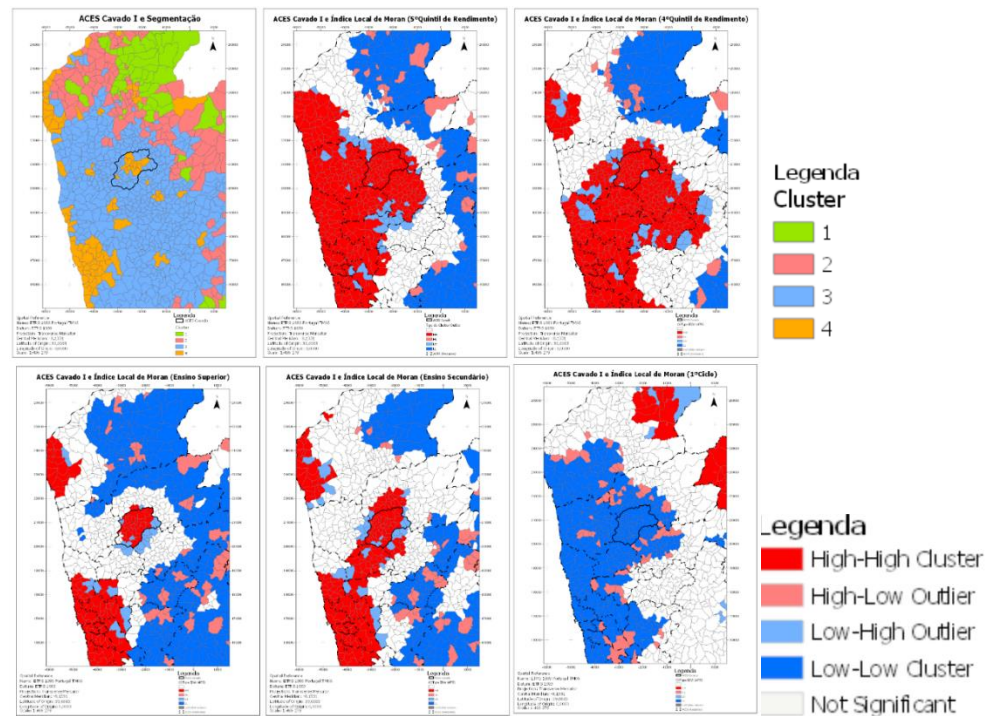


Figura 3.29 - ACES Cavado I com resultados da Segmentação e índices locais de Moran a alguns atributos

Uma vez que também foram utilizados dois atributos que podemos considerar comportamentais, o Gasto em Álcool e Gasto em Tabaco, figura 3.30, e graças ao índice local de Moran aplicado a estes atributos, podemos analisar estes por ACES. Podemos reparar que nas zonas onde os valores em proporção são mais elevados, são zonas que se encontram a azul forte, ou seja, os “Low-Low Cluster”, significa que são áreas que não existem elevados gastos neste dois tipos de substâncias, o que podemos, em certa parte, associar a hábitos mais saudáveis nesta zona. Em contra partida, na zona mais interior, existe um elevado número de “High-High Cluster” quando se considera o gasto em Tabaco, e no distrito de Lisboa e Setúbal, um elevado número de “High-High Cluster” referente ao gasto em Álcool, o que em certa parte, é refletido nos resultados demonstrados na figura 3.28 (página 65), pois são as zonas em que a proporção de utentes diabéticos inscritos que tem um resultado de HgbA1c igual ou inferior a 8% é menor, o que poderá estar aliado a maus hábitos comportamentais da população.

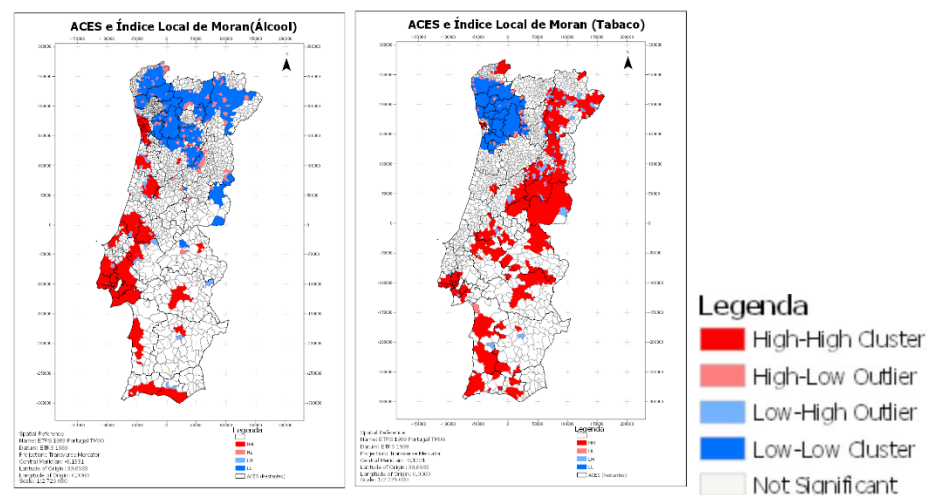


Figura 3.30 - Localização de ACES e índice local de Moran dos atributos "Gasto em Álcool" e "Gasto em Tabaco"

Resumindo, considerando estes dados relativo à diabetes, e todos os atributos e resultados da segmentação, podemos afirmar que os indivíduos das freguesias pertencentes ao *cluster* 3 e 4 apresentam mais indícios de ter um melhor controlo na sua doença, fazendo os devidos exames, seja por possuírem melhores níveis de rendimento, escolaridade e pertencerem a uma faixa etária mais baixa, ao contrário dos indivíduos pertencentes ao *cluster* 1 e 2, que são aqueles que se encontram nas zonas onde os valores das proporções eram os mais reduzidos, talvez por serem caracterizadas por uma população mais idosa, com menos posses e com um menor nível de escolaridade.



## 4. Produtos de Informação

Neste capítulo será detalhado quais os dados utilizados na criação dos produtos de informação, sendo depois apresentada a metodologia para cada um destes. No final é apresentado o *layout* inicial destes bem como os endereços em que podem ser consultados.

### 4.1 Metadados

Os dados utilizados neste projeto são maioritariamente de fontes públicas *online*, sendo que na sua maioria são dados que apresentam uma geometria do tipo polígono ou ponto devidamente georreferenciados e dados alfanuméricos organizados na forma tabular em folhas Excel.

Os dados sociodemográficos foram retirados do INE [11]. Estes, apresentados na Anexo IV, incluem maioritariamente o número de instituições de saúde, de profissionais de saúde, e toda a informação relevante da área da saúde presente nesta base de dados, sendo que se encontram no formato tabular (Excel) e com resolução espacial ao nível do município. Para além destes dados, foram também retirados os principais dados alfanuméricos e geográficos acerca dos censos de 2011. Estes estão inseridos numa camada de informação denominada por Base Geográfica de Referenciação de Informação (BGRI), que, como já referido, é um ficheiro que contém polígonos, e o seu sistema de coordenadas é o *ETRS 1989 TM06-Portugal*. O detalhe mais fino deste ficheiro é a subsecção estatística.

Os dados relativos à área da Saúde foram praticamente todos retirados da mesma fonte, uma base de dados do SNS, a Transparência [43]. Estes são do tipo pontual que se encontram no sistema de referência WGS 1984. Esta fornece dados relativo a temas sobre o SNS de Portugal, como o Acesso, Eficiência, Qualidade e Saúde dos Portugueses.

Na maioria, estes dados estão organizados ao nível de Agrupamentos de Centro de Saúde (ACES) ou Instituições Hospitalares, o nível mais fino de informação disponibilizado. Os dados usados estão caracterizados no Anexo IV, relativamente aos seus metadados.

Neste caso, todos os dados provenientes desta base de dados são camadas de informação do tipo ponto, umas ao nível dos ACES, outras ao nível das Instituições Hospitalares.

A figura 4.1 apresenta um mapa com a localização dos ACES em Portugal continental. As localizações pontuais são referentes à sede de cada ACES e no total são 55 sedes. É também apresentado um mapa das Instituições Hospitalares em Portugal continental. Existem cerca de 49 entidades hospitalares, desde Centros Hospitalares, Hospitais, Institutos e Unidades Locais de Saúde.

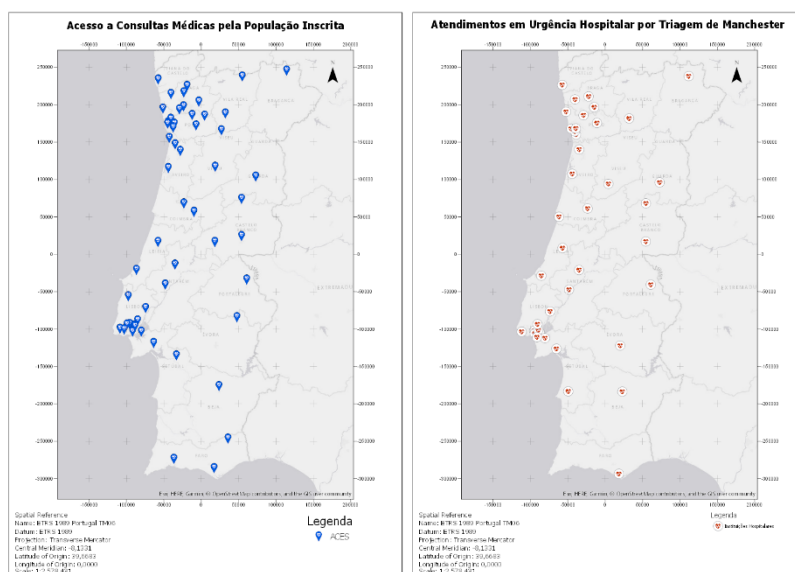


Figura 4.1 - Exemplo da localização dos dados com ACES como resolução espacial (Mapa de Acesso a Consultas Médicas pela População Inscrita) e Exemplo da localização dos dados com Instituições Hospitalares como resolução espacial (Mapa de Atendimentos em Urgência Hospitalar por Triagem de Manchester)

Os dados de localização de instituições de saúde, desde Hospitais, Centros de Saúde, Farmácias e Parafarmácias, foram retirados da base de dados denominada Geodados [45]. Estes são camadas de informação do tipo pontual em que o seu sistema de coordenadas é o WGS 1984 e a sua resolução espacial é a localização de cada instituição. O Geodados é uma plataforma disponibilizada pela Câmara Municipal de Lisboa que tem como objetivo a potenciação da utilização de informação produzida pelo município ao disponibilizar um conjunto de dados geográficos sobre Lisboa.

Na figura 4.2 encontram-se 3 mapas, um respeitante à localização dos hospitais públicos, outro à localização dos centros de saúde e outro à localização de farmácias e parafarmácias em Lisboa. No total existem cerca de 13 hospitais públicos e 107 centros de saúde. Existem cerca de 267 farmácias e parafarmácias distribuídas no concelho de Lisboa.

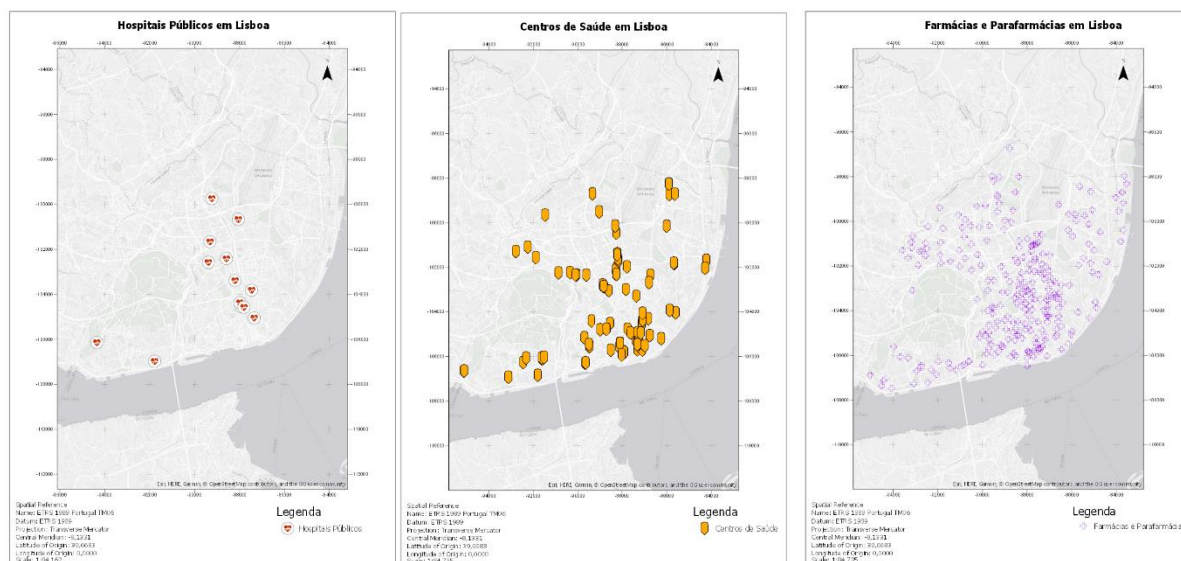


Figura 4.2 - Mapas respeitantes à localização de Hospitais Públicos, Centros de Saúde e Farmácias e Parafarmácias no concelho de Lisboa.

Novamente foi utilizada ferramenta *Enrich Layer*, de forma a enriquecer a camada de informação com os seguintes dados:

- Gastos em Produtos Médicos Total e per Capita – O gasto, em euros, abrange medicamentos, próteses, aparelhos e equipamentos médicos e outros produtos relacionados com a saúde adquiridos por indivíduos ou famílias, com ou sem receita médica, geralmente em farmácias ou fornecedores de equipamentos médicos. Estes são destinados a consumo fora de instituições de saúde [46].

- Poder de Compra Total e Per Capita – Consiste nos rendimentos líquidos, em euros, provenientes de emprego e ativos, após a dedução de impostos e contribuições sociais obrigatórias, pensões, benefícios de desemprego, pagamentos de benefícios e outros pagamentos de transferências nacionais [46].

Também foi utilizada a CAOP 2016, disponibilizada pela Direção Geral do Território [47]. Esta é uma camada de informação de geometria do tipo polígono, contendo a delimitação das áreas administrativas de Portugal, sendo o seu sistema de coordenadas o *ETRS 1989 Portugal TM06* e o seu detalhe mais fino é a freguesia.

Uma vez que também se pretende a criação de um *Operations Dashboard* com o resultado da segmentação, a camada de informação gerada anteriormente para a segmentação será utilizada.

## 4.2 Metodologia

Todos os dados foram numa primeira fase trabalhados em desktop no software *ArcGIS Pro*, sendo que para construir qualquer um dos produtos foi necessário serem publicados como serviço *web* para a plataforma *ArcGIS Online* com o objetivo de os adicionar a um mapa web e partilhá-los como um produto de informação.

### 4.2.1 *Operations Dashboard* de monitorização dos cuidados de saúde a nível de ARS e ACES

Neste *dashboard* foram incluídos dados da base de dados da Transparência, e também foi utilizada a CAOP. No Anexo V, encontram-se todos os dados utilizados e uma breve descrição destes.

Uma vez que parte dos dados se encontrava distribuídos por ARS, foi necessário também criar uma camada de informação de polígonos relativa a estas áreas geográficas. O ficheiro de partida foi o relativo a ACES, e sabendo a que ARS cada ACES pertence, foi selecionado o conjunto de ACES pertencente a uma ARS e aplicada a ferramenta *Dissolve*. Concluído o processo, este deu origem a uma camada de informação de polígonos representada na figura 4.3.

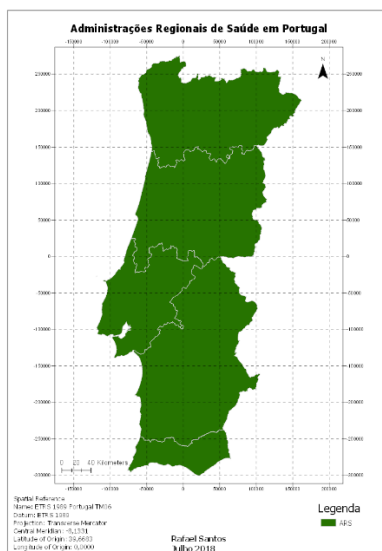


Figura 4.3 – Mapa com as Áreas Regionais de Saúde (ARS) em Portugal

Após a criação destas duas camadas de informação, o passo seguinte foi relacionar toda a informação proveniente dos dados com estas. Uma vez que a informação de todos os ficheiros se encontrava dividida por vários anos e vários ACES ou ARS, foi tomada a decisão de selecionar o ano completo mais recente, sendo que foi o ano de 2017. O *workflow* desta etapa encontra-se na figura 4.4.

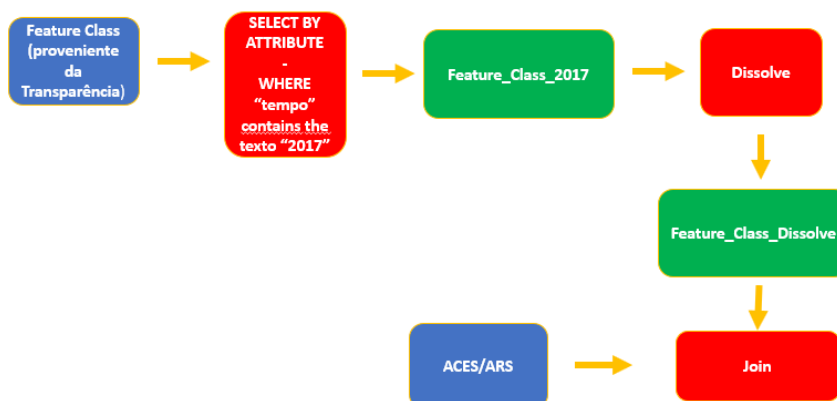


Figura 4.4 - Workflow da adição dos dados para a feature ACES/ARS

Uma vez que todos os dados provenientes da base de dados da Transparência são camadas de informação de pontos, recorreu-se primeiro à ferramenta de seleção por atributos e utilizou-se a condição *contains the text* igual a 2017, pois o campo que continha a informação sobre o ano era do tipo *text*. Na camada de informação criada após a utilização dessa ferramenta foi aplicada o *Dissolve*. Neste caso, o *Dissolve Field* foi o campo que continha o nome do ACES ou ARS, e a estatística aplicada aos campos da *feature* dependia da sua natureza. Caso fosse uma proporção ou taxa era aplicada a média do atributo, nos restantes casos era aplicada a soma. Para finalizar era feito um *Join* na *feature* referente aos ACES/ARS com informação proveniente da do *Dissolve* e um *Export Features* para salvar o acréscimo destas informações na *feature* dando origem a duas camadas de informação, ACES\_Transparência e ARS\_Final, sendo que em cada uma consta a localização de cada ACES ou ARS respetivamente.

Em seguida, para incorporar dados sociodemográficos neste *dashboard*, foi efetuado um *Spatial Join* da *feature* do ACES\_Transparência, com a da BGRI. É uma relação de um-para-muitos uma vez que cada ACES tem várias freguesias. Esta ferramenta deu origem a uma camada de informação

ACES\_Transparencia\_Join\_BGRI. Uma vez que queremos a informação por ACES, foi necessário efetuar novamente um *Dissolve*, sendo que o *Dissolve Field* é o campo que contém o nome do ACES, e a estatística aplicada aos campos foi a soma. A camada de informação proveniente deste processo é a ACES\_Transparencia\_BGRI. Os dados dos censos que interessam para este *dashboard* foram apenas o número total de indivíduos residentes, por sexo e por faixa etária, sendo que os restantes campos foram eliminados.

Uma vez que é também do interesse ter a informação relativa a ACES por ARS foi necessário efetuar novamente um *Spatial Join*. À camada de informação ACES\_Transparencia\_BGRI, foi adicionada o nome da ARS que cada ACES pertence, através da *feature* ARS\_Transparencia. A relação foi de um-para-um, pois cada ACES tem apenas uma ARS, dando origem a uma camada de informação denominada por ACES\_Final.

Ao obter as duas camadas de informação desejadas, como referido no início do capítulo, foi efetuado um *Publish Web Layer* com estas, ou seja, foi feito o *upload* das camadas de informação para o servidor do *ArcGIS Online*. Após aceder ao *ArcGIS Online*, foi criado um mapa *web* para adicionar as duas camadas e proceder à criação do *dashboard*.

#### 4.2.2 *Operations Dashboard* de monitorização/gestão proativa dos cuidados de saúde a nível de Municípios e Distritos

Neste *Operational Dashboard* foram incluídos maioritariamente dados da base de dados do INE, sendo estes os seguintes: Centros de saúde por localização geográfica e tipo de serviço, Enfermeiras/os por Local de trabalho e Sexo, Médicas/os por Local de trabalho e Sexo, Farmacêuticas/os por Local de trabalho e Sexo, Farmácias e postos farmacêuticos móveis por localização geográfica e Tipo de unidade local de farmácia, Hospitais por Localização geográfica e Natureza institucional, Atendimentos em serviço de urgência nos hospitais por Localização geográfica, Atendimentos no serviço de urgência básica – SSUB ou no serviço de atendimento permanente ou prolongado nos centros de saúde, Consultas médicas na unidade de consulta externa dos hospitais por Localização geográfica e Especialidade da consulta, Pessoal ao serviço nos hospitais públicos de acesso universal e hospitais em parceria público-privada por localização geográfica e tipo de pessoal ao serviço, Pessoal ao serviço nos centros de saúde por Localização geográfica e tipo de pessoal ao serviço, Taxa bruta de mortalidade por Local de residência, Taxa bruta de natalidade por Local de residência, Óbitos por Local de residência, Sexo, Grupo etário e Causas de morte, Partos nos hospitais públicos de acesso universal e hospitais em parceria público-privada por Localização geográfica e Tipo de parto, Interrupções voluntárias da gravidez legalmente efetuadas em estabelecimentos de saúde por Local de residência e Quota de mercado do medicamento genérico por localização geográfica.

Foram também utilizados alguns dados provenientes da ferramenta *Enrich Layer* da ESRI, nomeadamente o Gasto em Produtos Médicos Total e per Capita, Poder de Compra Total e per Capita, e também foi utilizada a CAOP.

Uma vez que o objetivo deste *Operational Dashboard* é ter informação organizada a nível de municípios e distritos foi necessário criar duas camadas de informação que tivessem esta informação. Para isso recorreu-se a camada de informação da CAOP e à ferramenta de geoprocessamento *Dissolve*, pois a camada de informação encontra-se no seu grão mais fino, polígonos referentes à localização das freguesias em Portugal.

No caso da camada de informação com as localizações dos municípios, ao aplicar o *Dissolve* na camada de informação CAOP, o *Dissolve Field* é o campo que contém o nome de cada município, dando

origem a uma camada de informação de polígonos denominada por “Municípios”, figura 4.5. Em Portugal continental existem 278 municípios, sendo que existem mais 11 na Madeira e 19 nos Açores, mas neste caso apenas teremos em conta os que se encontram no continente.

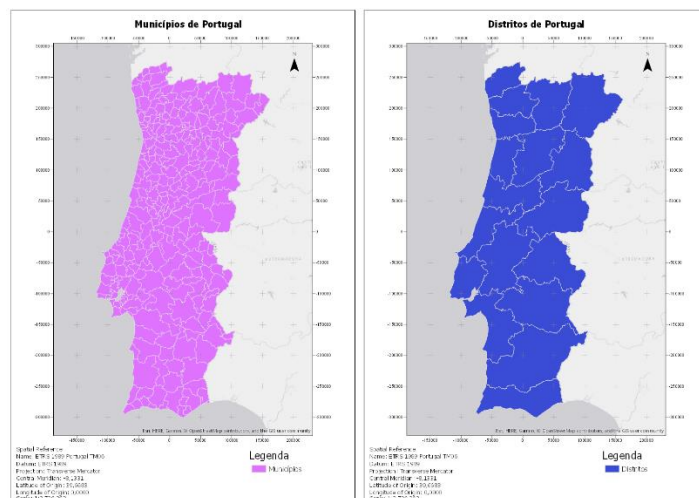


Figura 4.5 - Mapas com os Municípios (esquerda) e Distritos (direita) de Portugal

Para criarmos a camada de informação com as localizações de cada distrito, o processo é idêntico ao anterior, mas neste caso o *Dissolve Field* é o campo que contém o nome de cada distrito. A camada de informação de polígonos criada a partir deste processo denominada por “Distritos” encontra-se à direita na figura 4.5. Em Portugal existem 18 distritos.

Em seguida, para incorporar dados sociodemográficos, recorreu-se à ferramenta de geoprocessamento *Join*, para adicionar informação à camada de informação “Municípios”, sendo que o atributo de cada tabela utilizado para efetuar o *Join* era o nome do município. Após este passo ter sido efetuado com todas as tabelas com informação proveniente do INE, recorreu-se à ferramenta de geoprocessamento *Enrich Layer* para adicionar a informação. No final obteve-se uma camada de informação de polígonos denominada por “Municípios\_Informação” com as localizações de cada município e toda a informação sociodemográfica.

Uma vez que esta última camada continha um atributo que correspondia ao distrito ao qual cada município pertence bastou apenas recorrer à ferramenta de geoprocessamento *Join* para adicionar toda a informação à camada Distritos, dando assim origem à camada de informação “Distritos\_Informacao”

Como no subcapítulo anterior, foi efetuado um *Publish Web Layer* com as duas camadas de informação, ou seja, foi feito o *upload* destas para o servidor do *ArcGIS Online*. Após aceder ao *ArcGIS Online*, foi criado um mapa *web* para adicionar as duas camadas e proceder à criação do *dashboard*.

### 4.2.3 Operations Dashboard com o resultado da segmentação e dados da Diabetes

Este produto de informação utilizou os resultados provenientes da segmentação, ou seja, a camada de informação “Segmentacao”, sendo que esta tinha toda a informação relativa aos atributos utilizados e a sua resolução espacial é ao nível da freguesia. Também foram utilizados os dados relativos a diabetes provenientes da base de dados da Transparência e os dados “Óbitos (N.º) por Local de residência (NUTS - 2013), Sexo, Grupo etário e Causa de morte (Lista sucinta europeia) (2016)” e a camada de informação BGRI provenientes do INE.

Uma vez que tanto a camada de informação “Segmentação” e a “BGRI” são camadas ao nível da freguesia, ambas com o código DICOFRE como atributo, foi apenas necessário recorrer à ferramenta de geoprocessamento *Join* para juntar toda a informação de ambas num só ficheiro denominado por “Segmentação\_BGRI”, sendo esta uma camada de informação de polígonos com a informação proveniente das duas camadas mencionadas anteriormente.

A informação dos óbitos devido a diabetes encontrava-se ao nível do município, para tal recorreu-se novamente ao *Join* para acoplar esta informação à camada de informação “Segmentação\_BGRI”, tendo o atributo “Concelho” como atributo comum.

Para incorporar os dados da diabetes foi necessário utilizar uma camada de informação criada para o *dashboard* ao nível de ACES, denominada de “ACES\_Transparencia\_Join\_BGRI”, sendo esta uma camada de informação de polígonos que tinha qual o ACES correspondente a cada freguesia. Foi utilizado novamente o *Join* para efetuar a adição da informação da diabetes à camada “Segmentacao\_BGRI”, tendo o atributo DICOFRE em comum, para então dar origem à camada de informação final “Segmentacao\_Dashboard”. Esta é uma camada de informação de polígonos com toda a informação referente à segmentação, dados dos censos de 2011, número de óbitos devido a diabetes e dados sobre o programa de atividade da diabetes proveniente da Transparência.

#### 4.2.4 Visualizador com áreas de influência de Hospitais em Lisboa

Para a criação deste visualizador geográfico foram necessárias apenas os dados referentes às localizações das instituições da base de dados do Geodados da Câmara Municipal de Lisboa, nomeadamente Farmácias e Parafarmácias, Hospitais Públicos e Centros de Saúde, e a CAOP.

O primeiro objetivo deste visualizador foi a criação de áreas de influência dos Hospitais Públicos em Lisboa. Uma área de influência é um polígono que define qual a área que pode ser alcançada a partir de uma determinada localização (*facility*) até um determinado custo de trajeto, seja este uma medida de distância ou tempo. Para esse efeito, recorreu-se à ferramenta de geoprocessamento *Generate Service Areas*. Como *Facilities* utilizou-se a camada de informação dos Hospitais Públicos, os *Break Values* considerados foram de 5, 10 e 15, o *Break Units* foi *Minutes*, a *Travel Direction* foi *Towards Facility*., a *Impedance* foi *Drive Time* e o *Polygon Overlap Type* foi *Disks*. Ou seja, a ferramenta gerou áreas nas quais um indivíduo demora respetivamente 5, 10 e 15 a chegar até ao hospital de carro.

É importante perceber qual o impacto do trânsito em Lisboa, uma vez que sendo a capital do país e das zonas com mais população, é natural que haja horas de grande afluência de carros nas estradas alterando o tempo que um utente demora a alcançar uma unidade de saúde ou o tempo que uma ambulância consome para chegar a uma localização. Tendo isso em conta, foram geradas duas camadas de informação com áreas de influência com os parâmetros acima mencionados, uma para uma Segunda Feira às 9h da manhã, considerada hora de trânsito, tendo sido escolhido para a simulação o dia 07/05/2018, e outra para um Sábado à mesma hora, dia 05/05/2018, considerada hora de tráfego “normal”.

Uma vez que a camada dos Hospitais Públicos é constituída por 13 entradas, ou seja, 13 Hospitais Públicos, figura 4.2, a ferramenta irá gerar uma camada de informação com 13\* 3 áreas, ou seja, 39 entradas.

Primeiramente foi gerada as áreas de influência para a hora de trânsito, estando um exemplo à esquerda na figura 4.6. Como é de esperar, a área correspondente a 5 minutos é reduzida, sendo que na maioria as áreas são referentes a 15 minutos de carro. Ainda assim, verifica-se que existe um ou mais



hospitais públicos no concelho de Lisboa que é alcançável em 15 minutos a partir da margem sul do Tejo.

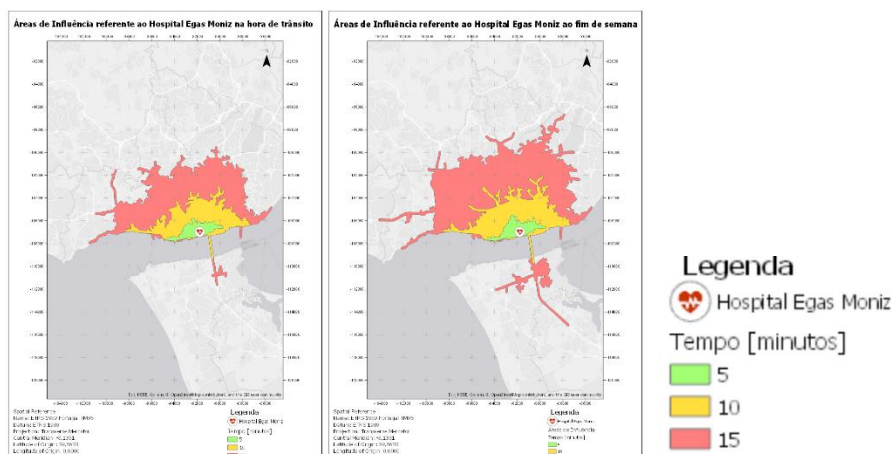


Figura 4.6 - Mapa com Áreas de Influência referente ao Hospital Egas Moniz na hora de trânsito (esquerda) e ao fim de semana (direita)

Quanto às áreas de influência referentes ao horário de sábado de manhã, exemplo à direita na figura 4.6, como seria de esperar existe uma maior área de influência, sendo que claramente áreas referentes ao tempo de 15 minutos no concelho de Lisboa na figura anterior, ao fim de semana tornam-se áreas referentes ao tempo igual a 10 minutos, ou seja, áreas que a uma segunda de manhã demorariam cerca de 15 minutos a chegar a um hospital, a um sábado de manhã à mesma hora demorariam cerca de 10 minutos.

A etapa seguinte de construção deste visualizador foi gerar duas camadas de informação, uma para cada *service area* [48] (trânsito e fim de semana), que representasse as freguesias que são abrangidas pela *service area* em pelo menos 20% da sua área com cada um dos tempos escolhidos (5,10,15).

Pretendia-se uma camada de informação para cada entrada da camada das áreas de influência (*service area*), ou seja, 39 camadas de polígonos, cada uma referente a um tempo para cada hospital. Por exemplo, a *facility* 1, referente ao Hospital Egas Moniz, irá ter 3 camadas, uma para as freguesias que são alcançadas em 5 minutos, uma para 10 minutos e outra para 15 minutos. O primeiro passo foi criar um modelo de geoprocessamento, figura 4.7, que permitisse automatizar o processo.

O modelo efetua o seguinte cálculo: para cada área de influência na camada original é efetuada uma intersecção espacial com a camada referente às freguesias de Lisboa, originando uma camada de informação com as freguesias que são alcançadas a partir de um hospital a um determinado tempo. Em seguida é executado o *script* que irá dar origem ao ficheiro final desejado.

Este script (anexo IV) faz a verificação se pelo menos 20% da área das freguesias são abrangidas pela área de influência, não sendo, não são contabilizadas. Caso nenhuma freguesia fosse abrangida em pelo menos 20%, a freguesia na qual a *facility* está localizada é automaticamente selecionada para permanecer na camada de informação.

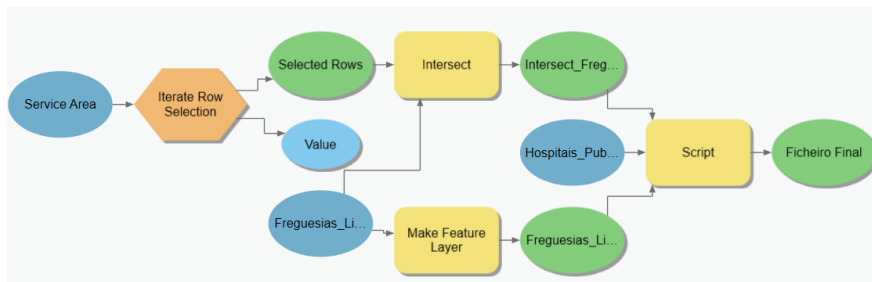


Figura 4.7 - Modelo de geoprocessamento

Através da camada CAOP e de uma seleção de atributos foram seleccionadas as freguesias que fossem do concelho de Lisboa.

Na figura 4.8 está o exemplo de 3 saídas deste modelo, uma para as freguesias nas quais é possível chegar ao Hospital Egas Moniz em 5 minutos, outra em 10 minutos e outra em 15 minutos, com a verificação da condição de que pelo menos 20% da freguesia está incluída na área de influência.

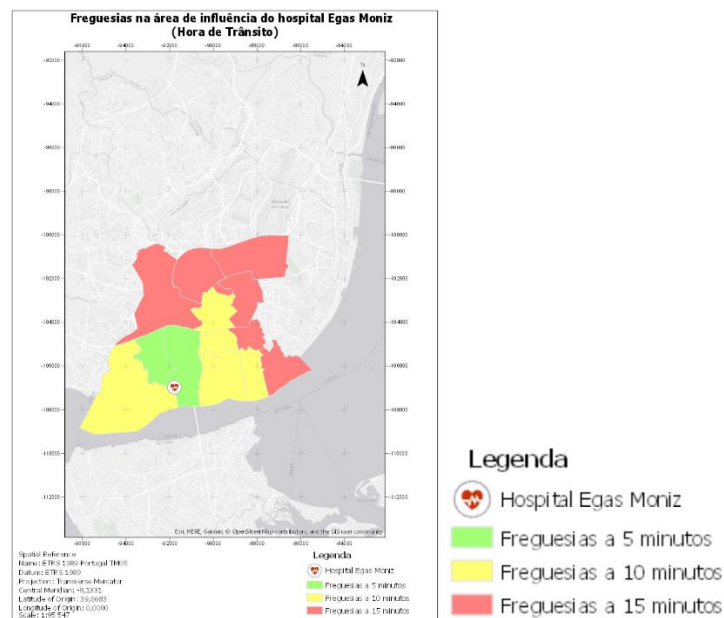


Figura 4.8 - Mapa com as Freguesias na área de influência do hospital Egas Moniz (Hora de Trânsito)

Após este passo, foi criada uma nova camada denominada por Freguesias\_Transito através da ferramenta *Create Feature Class*. Nesta foi utilizada a ferramenta *Append*, com o intuito de adicionar todas as camadas de informação criadas pelo modelo de geoprocessamento, para obter uma camada de informação que tivesse todas as freguesias alcançadas num determinado tempo, numa determinada *facility*.

Todo este processo foi repetido, mas para a camada de informação referente às áreas de influência durante o fim de semana.

De forma reduzir o número de camadas de informação, em ambas as camadas com as áreas de influência nas diferentes horas de trânsito foi adicionado um atributo denominado por Type\_Traffic, preenchido de acordo com a hora de trânsito escolhida, sendo “Hora de Trânsito” para as *features* com hora de trânsito e “Hora Fora de Trânsito” para a *feature* que foi gerada tendo em conta o horário de fim de semana. Para terminar, recorreu-se novamente ao *Create Feature Class* para dar origem a um ficheiro

final denominado por “Áreas de Influência”, que contem todas as freguesias dentro de cada área de influência geradas anteriormente.

Uma vez que se pretende também dar ao utilizador a oportunidade de verificar qual o Centro de Saúde ou farmácia mais próxima da localização que escolhe através do visualizador, foi necessário adicionar também as camadas de informação correspondentes às localizações destes elementos.

Como anteriormente, foi efetuado um *Publish Web Layer* mas neste caso com quatro camadas de informação: áreas de serviço, freguesias, localizações dos Hospitais Públicos, Centros de Saúde e localização de Farmácias e Parafarmácias. No *ArcGIS Online*, foi criado um mapa *web* para adicionar estas camadas e proceder à criação do visualizador.

Na criação do visualizador é possível incluir *widgets*, que acabam por ser ferramentas possíveis de utilizar dentro deste produto. Neste caso foram selecionados 3 *widgets* a disponibilizar: **Filtro** (permite filtrar qual a área de influência de um hospital e quais as freguesias que estão no seu alcance num determinado tempo), **Perto de Mim** (mostra quais as farmácias, parafarmácias, centros de saúde e hospitais públicos mais próximos, dentro de um *buffer* definido pelo utilizador, de um endereço introduzido na barra de endereços ou de um ponto definido no mapa) e **Sobre** (permite a inclusão de texto informativo, foi utilizado para fornecer instruções sobre como utilizar e tirar o maior partido deste produto).

## 4.3 Resultados

Como referido nos objetivos deste projeto, foram criados quatro produtos de informação para demonstrar várias vertentes da área da saúde em Portugal.

Quanto ao *Operational Dashboards* de monitorização dos cuidados de saúde a nível de ARS e ACES, figura 4.9, este demonstra indicadores que avaliam a performance dos ACES e das ARS, sendo que também apresenta todos os dados relativos às atividades dos programas de diabetes, Hipertensão, Vacinação promovidos pelo estado. É possível selecionar estas áreas através da sua designação ou então através da Taxa de Utilização Global de Consulta Médica. Este produto encontra-se disponível no endereço: <https://arcg.is/0mr8qX>

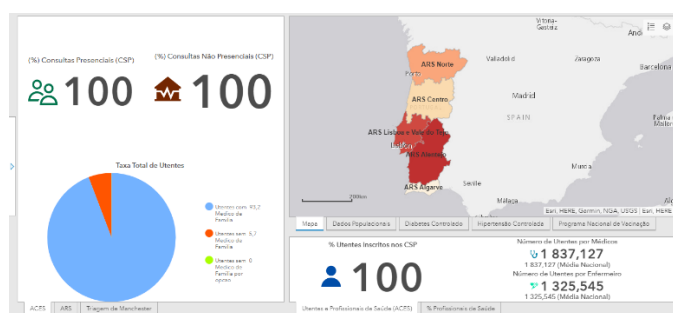


Figura 4.9 - *Operational Dashboards* de monitorização dos cuidados de saúde a nível de ARS e ACES

O *Operational Dashboard* de monitorização/gestão proativa dos cuidados de saúde a nível de Municípios, figura 4.10, apresenta todos os indicadores gerais do estado do setor da saúde em Portugal, sejam dados de pessoal ao serviço, o número das variadas instituições de saúde, a economia na área da saúde, e características da população na área selecionada. Este permite visualizar os dados a nível de município ou distrito, selecionado a seta do lado esquerdo. O endereço em que este se encontra disponível é: <https://arcg.is/1a0rTW>



Figura 4.10 - Operational Dashboard de monitorização/gestão proativa dos cuidados de saúde a nível de Municípios

Sobre o *Operational Dashboards* que demonstra o resultado da segmentação e de certa forma todos os dados públicos referentes à diabetes, presente na figura 4.11, tem como objetivo mostrar todos os indicadores dos atributos utilizados na segmentação, sendo possível fazer filtros através de Distritos, Concelhos ou até mesmo por *Clusters*. Estão incluídos também os dados utilizados na análise da diabetes. Este produto encontra-se disponível no endereço: <https://arcg.is/0jm54>

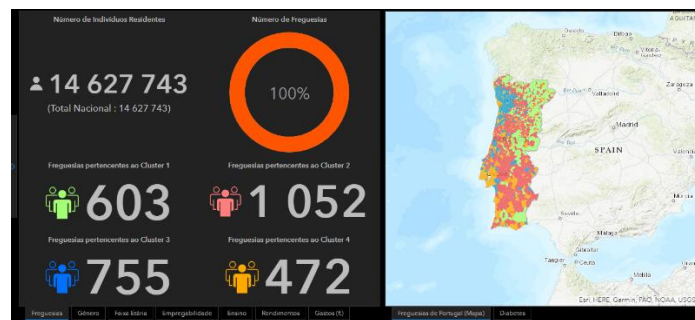


Figura 4.11 - Operational Dashboards que demonstra o resultado da segmentação e contém dados da diabetes

O *Visualizador* com áreas de influência de Hospitais em Lisboa, figura 4.12, permite verificar quais são essas mesmas áreas caso seja uma hora que exista trânsito, ou não, e quais as freguesias que se encontram nessas mesmas áreas. Também permite ao utilizador colocar um endereço e verificar qual o hospital, centro de saúde e farmácia mais perto. Ao iniciar a aplicação, esta terá uma mensagem com as instruções necessárias para tirar máximo partido deste mesmo produto. O endereço em que este se encontra disponível é o: <https://esriportugal.maps.arcgis.com/apps/webappviewer/index.html?id=171e187545204baba1cb0941dd4c8e42>

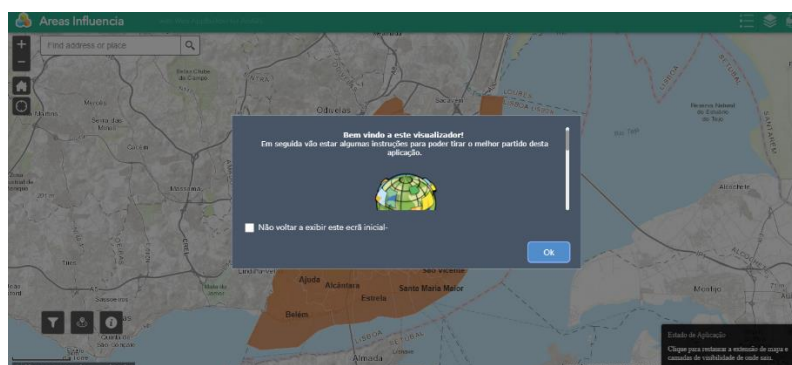


Figura 4.12 - Visualizador com áreas de influência de Hospitais em Lisboa

## 5. Conclusões Finais e Trabalhos Futuros

O principal objetivo deste projeto era tentar elaborar uma metodologia, que pudesse compreender alguns fatores socioeconómicos que pudessem estar na origem de problemas associados à saúde, sendo que neste caso foram utilizados dados relativos à diabetes.

Com este projeto foi possível verificar que aliar uma ferramenta não-espacial, neste caso um algoritmo de *Clustering*, com ferramentas de análise espacial, neste caso o Índice Local de Moran, é bastante proveitoso pois ambos os métodos são independentes mas complementam-se nos seus resultados, permitindo uma análise mais completa da caracterização da população relativamente aos seus hábitos de consumo, características socioeconómicas e inferir sobre a qualidade da saúde dessas populações.

Através do resultado da segmentação, os quatros grupos gerados, com as suas devidas características, acabam por dar uma noção de como as freguesias se diferenciam entre si. Por outro lado, a análise ao Índice Local de Moran, demonstra quais os atributos que contribuem para tal diferenciação, e uma vez que os *clusters* gerados na segmentação são o conjunto dos atributos, através deste índice é possível verificar quais são realmente os atributos determinantes em cada freguesia. Por exemplo, de uma forma geral, o *cluster* 4 é caracterizado por ter indivíduos jovem adultos empregados com níveis altos de rendimento e escolaridade, mas existem freguesias pertencentes a este *cluster* que podem ter valores altos de pensionistas e reformados, e essa característica não pertencer a este *cluster*.

Com os atributos utilizados, Portugal não demonstra uma grande versatilidade em termos de características populacionais, tanto que de fato o número de *clusters* ideal sugerido pelo teste de estatística Pseudo-F era 2, e espacialmente o continente acabava por ser encontrar dividido numa zona interior com determinadas características e numa zona litoral. Em termos de atributos, por norma existem mais indivíduos do género Feminino do que Masculino, mais indivíduos com faixa etária de 24 aos 64 anos de idade, mais indivíduos com o 1º ciclo completo, apenas em termos de ensino e rendimento sendo possível visualizar alguma diferença através dos gráficos presentes no subcapítulo 3.3.

Relativamente ao caso dos diabetes, com os resultados gerados através da segmentação e da aplicação dos Índices Locais de Moran foi possível verificar que os indivíduos do *cluster* 3 e 4 são aqueles que têm uma maior probabilidade de se deslocarem a uma instituição de saúde e tratar e controlar a sua doença, pois são aqueles que apresentam um melhor valor de proporções relativo aos dados da doença e apresentam um maior nível de rendimento e educação. No entanto, numa doença como a diabetes seria importante utilizar atributos que demonstrassem comportamentos dos indivíduos, seja hábitos alimentares, hábitos de exercício físico, e outros atributos que pudessem demonstrar a forma como o indivíduo leva a sua vida, um pouco à semelhança do exemplo demonstrado no capítulo teórico sobre a segmentação na área da saúde, subcapítulo 2.4, pois a diabetes é uma doença que se baseia muito no comportamento do indivíduo (hábitos alimentares e sedentarismo).

Informação relativa à área da saúde, em termos geoespaciais, é raramente disponibilizada ao público em geral. O portal Geosaúde [49] disponibilizado pelo serviço nacional de saúde é um exemplo de disponibilização de informação, contudo o acesso a dados que pudessem ser relevantes para este projeto foi bastante complicado.

Deve ainda salientar-se que a tecnologia ESRI foi fundamental e útil para o desenvolvimento deste projeto, fornecendo ferramentas que possibilitaram agilizar alguns procedimentos, nomeadamente a utilização do algoritmo de *Clustering* através do software ArcGIS em vez de recorrer a um script com

uma linguagem de programação. O processo da criação dos produtos de informação foi intuitivo e de compreensão fácil.

Após a realização deste projeto, e caso profissionais da área da saúde pretendam implementar as metodologias aqui descritas, recomenda-se que essas sejam sempre complementada com outros dados para além dos provenientes dos Censo 2011. Por exemplo, dados associados a hábitos de vida, dados de incidência de patologias, entre outros. A resolução espacial continua a ser uma grande limitação na utilização desta informação porque o nível de generalização da mesma, por vezes não permite compatibilizar estes dados com níveis de elevado detalhe espacial como por exemplo, a BGRI, ou até mesmo ao nível da freguesia. O acesso a dados relacionados com a área da saúde continua a ser ainda uma limitação grande para quem pretende trabalhar e disponibilizar informação da área da saúde, sendo, portanto, imperativo que as instituições publicas e privadas atuem em parceria na cedência de informação de modo a serem elaboradas bases de dados ao nível nacional que possibilitem estudos como este mas com maior completude e confiança nos dados que são usados.

## Referências Bibliográficas

- [1] F. Ritter, R. D. Rosa, and R. Flores, “Health situation assessment by primary care workers based on geographic information systems,” *Cad. Saude Publica*, vol. 29, no. 12, pp. 2523–2534, 2013.
- [2] G. M. Foody, “GIS : health applications,” vol. 5, pp. 691–695, 2006.
- [3] The Better Care Fund, “‘How to’ Guide: The BCF Technical Toolkit,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.
- [4] L. Anselin, “Local Indicators of Spatial Association—LISA,” *Geogr. Anal.*, vol. 27, no. 2, pp. 93–115, 1995.
- [5] P. Tsasis *et al.*, “Conceptualizing type 2 Diabetes and its management,” *J. Multidiscip. Healthc.*, vol. 9, pp. 133–136, 2016.
- [6] L. Rebekah Gospin, James P. Leu, and Joel Zonszein, Poretsky, *Principle of Diabetes Mellitus*. 2017.
- [7] International Diabetes Federation, “International Diabetes Federation - Facts & Figures.” [Online]. Available: <https://www.idf.org/aboutDiabetes/what-is-Diabetes/facts-figures.html>. [Accessed: 16-Jul-2018].
- [8] DGS, “Programa Nacional para Diabetes,” *Plano Nac. da Diabetes*, p. 20, 2017.
- [9] Sociedade Portuguesa de Diabetologia, *Diabetes: Fatos e Números – O Ano de 2015 – Relatório Anual do Observatório Nacional da Diabetes*. 2016.
- [10] Ministério da Saúde de Portugal, “Relatório Anual de Acesso a Cuidados de Saúde nos estabelecimentos do SNS e Entidades Convencionadas (2017),” 2017.
- [11] INE, “Instituto Nacional de Estatística.” [Online]. Available: <https://www.ine.pt/>. [Accessed: 01-Aug-2018].
- [12] K. Kurland and W. Gorr, *GIS Tutorial for Health*, 5th ed. ESRI Press, 2014.
- [13] G. Rushton, “Public Health, GIS, and Spatial Analytic Tools,” *Annu. Rev. Public Health*, vol. 24, no. 1, pp. 43–56, 2003.
- [14] C. I. J. Nykiforuk and L. M. Flaman, “Geographic Information Systems (GIS) for Health Promotion and Public Health: A Review,” *Health Promot. Pract.*, vol. 12, no. 1, pp. 63–73, 2011.
- [15] R. M. da Silva, H. R. de F. Pinto, S. G. F. Costa, and K. R. de F. Pinto, “Análise espacial da endemia hansenica no município de Bayeux (PB) mediante técnicas de SIG,” *Soc. Nat.*, vol. 24, no. 2, pp. 345–358, 2012.
- [16] J. Maantay, “Asthma and air pollution in the Bronx: Methodological and data considerations in using GIS for environmental justice and health research,” *Heal. Place*, vol. 13, no. 1, pp. 32–56, 2007.
- [17] V. Thomas, Nihal; Kapoor, Nitin; Velavan, Jachin; Senthil, *A Practical Guide to Diabetes Mellitus*. 2016.
- [18] Associação Protectora dos Diabetes de Portugal, “O que é a Diabetes?” [Online]. Available: <http://www.apdp.pt/Diabetes/a-pessoa-com-Diabetes/o-que-e-a-Diabetes>.

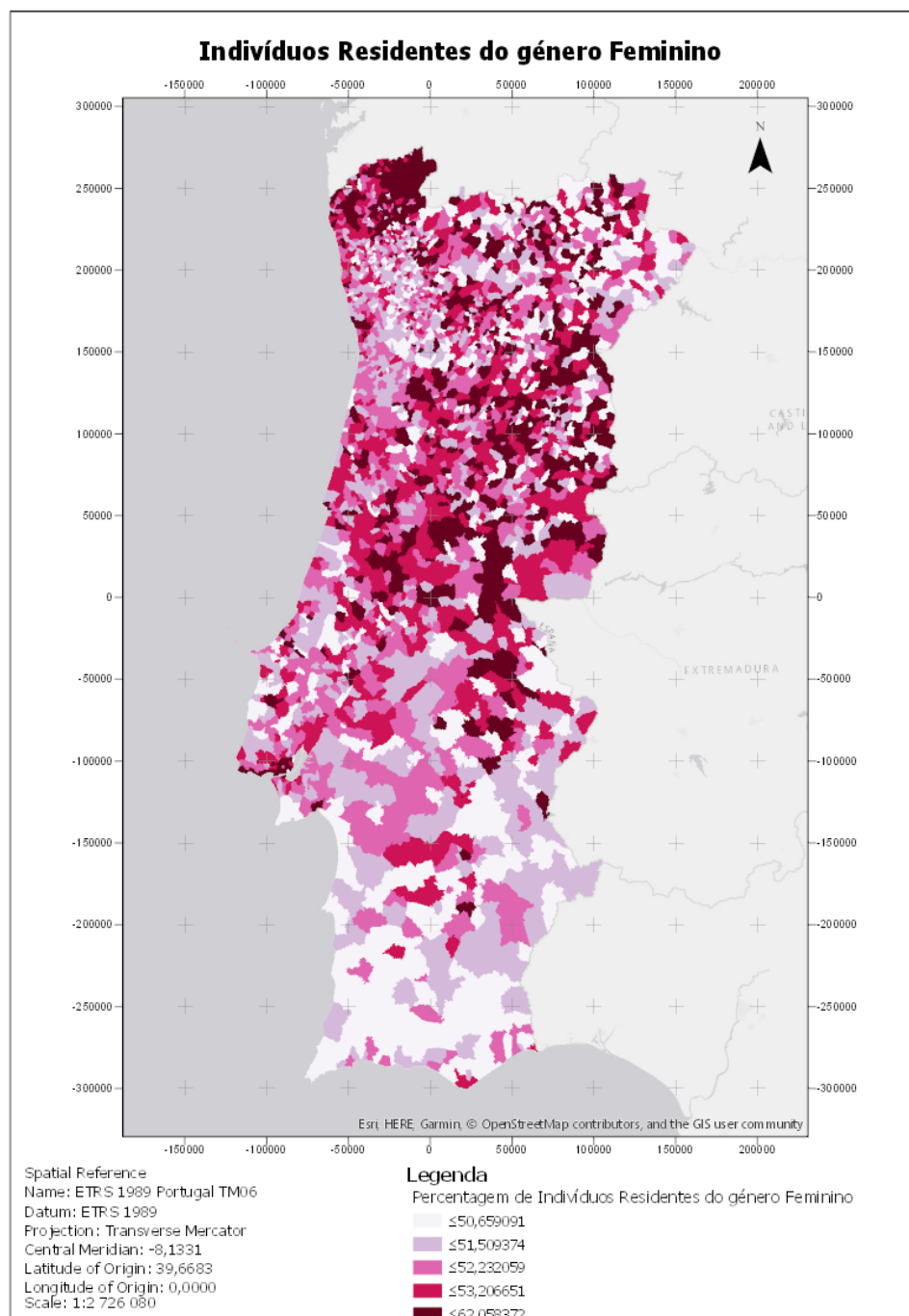


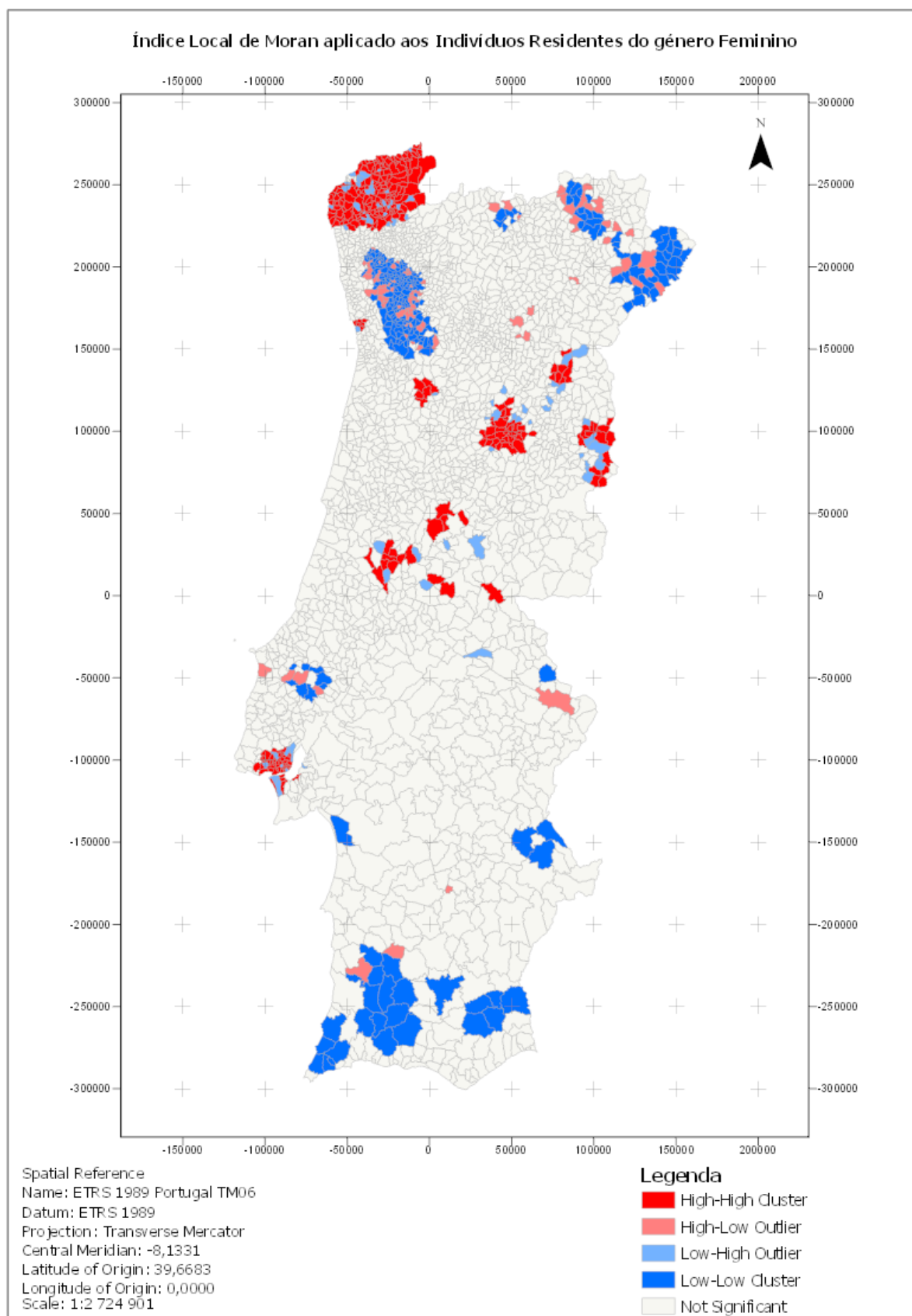
- [19] Direção Geral da Saúde, “Prescrição e Determinação da Hemoglobina Glicada A1c,” *Norma Da Direção Geral Da Saúde*, vol. 1, no. 1, p. 9, 2013.
- [20] E. M. Geraghty, T. Balsbaugh, J. Nuovo, and S. Tandon, “Using Geographic Information Systems (GIS) to Assess Outcome Disparities in Patients with Type 2 Diabetes and Hyperlipidemia,” *J. Am. Board Fam. Med.*, vol. 23, no. 1, pp. 88–96, 2010.
- [21] U. Samuelsson and O. Löfman, “Geographical mapping of type 1 Diabetes in children and adolescents in south east Sweden,” *J. Epidemiol. Community Health*, vol. 58, no. 5, pp. 388–392, 2004.
- [22] S. Mohan *et al.*, “UDAY: A comprehensive Diabetes and hypertension prevention and management program in India,” *BMJ Open*, vol. 8, no. 6, p. e015919, 2018.
- [23] A. B. Curtis, C. Kothari, R. Paul, and E. Connors, “Using GIS and secondary data to target Diabetes-related public health efforts,” *Public Health Rep.*, vol. 128, no. 3, pp. 212–220, 2013.
- [24] C. Sousa and A. Monteiro, “Diabetes Mellitus tipo 2 na área metropolitana do Porto - O risco de uma realidade amarga provocada também pelo (des)ordenamento do território,” vol. 2, no. I, pp. 103–111, 2016.
- [25] B. Williams, D. McVey, L. Davies, and E. MacGregor, “The Healthy Foundations Lifestage Segmentation: Research Report No.1: Creating the Segmentation using a quantitative survey of the general population of England.,” no. 1, pp. 1–312, 2011.
- [26] scikit-learn, “2.3 - Clustering,” 2017. [Online]. Available: <http://scikit-learn.org/stable/modules/clustering.html#k-means>. [Accessed: 02-Sep-2018].
- [27] S. Guido and A. Müller, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 2017.
- [28] ESRI, “Multivariate Clustering.” [Online]. Available: <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/multivariate-clustering.htm>. [Accessed: 05-Sep-2018].
- [29] A. Trevino, “Introduction to K-means Clustering.” [Online]. Available: <https://www.datascience.com/blog/k-means-clustering>. [Accessed: 24-Sep-2018].
- [30] J. Burkardt, “K-means Clustering.” [Online]. Available: [https://people.sc.fsu.edu/~jburkardt/classes/isc\\_2009/clustering\\_kmeans.pdf](https://people.sc.fsu.edu/~jburkardt/classes/isc_2009/clustering_kmeans.pdf). [Accessed: 25-Aug-2018].
- [31] NHS Tower Hamlets Clinical Commissioning Group, “Tower Hamlets CCG,” 2018. [Online]. Available: <http://www.towerhamletscg.nhs.uk/>.
- [32] Experian, “Mosaic,” 2018.
- [33] Change4Life, “Change4Life.” [Online]. Available: <https://www.nhs.uk/change4life>. [Accessed: 31-Aug-2018].
- [34] S. L. McLafferty, “GIS and Health Care,” *Annu. Rev. Public Health*, vol. 24, no. 1, pp. 25–42, 2003.
- [35] ESRI, “How Cluster and Outlier Analysis (Anselin Local Moran’s I) works,” 2018. [Online]. Available: <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-cluster-and-outlier-analysis-anselin-local-m.htm>. [Accessed: 13-Sep-2018].
- [36] ESRI, “Modeling spatial relationships,” 2018. [Online]. Available: <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/modeling-spatial->

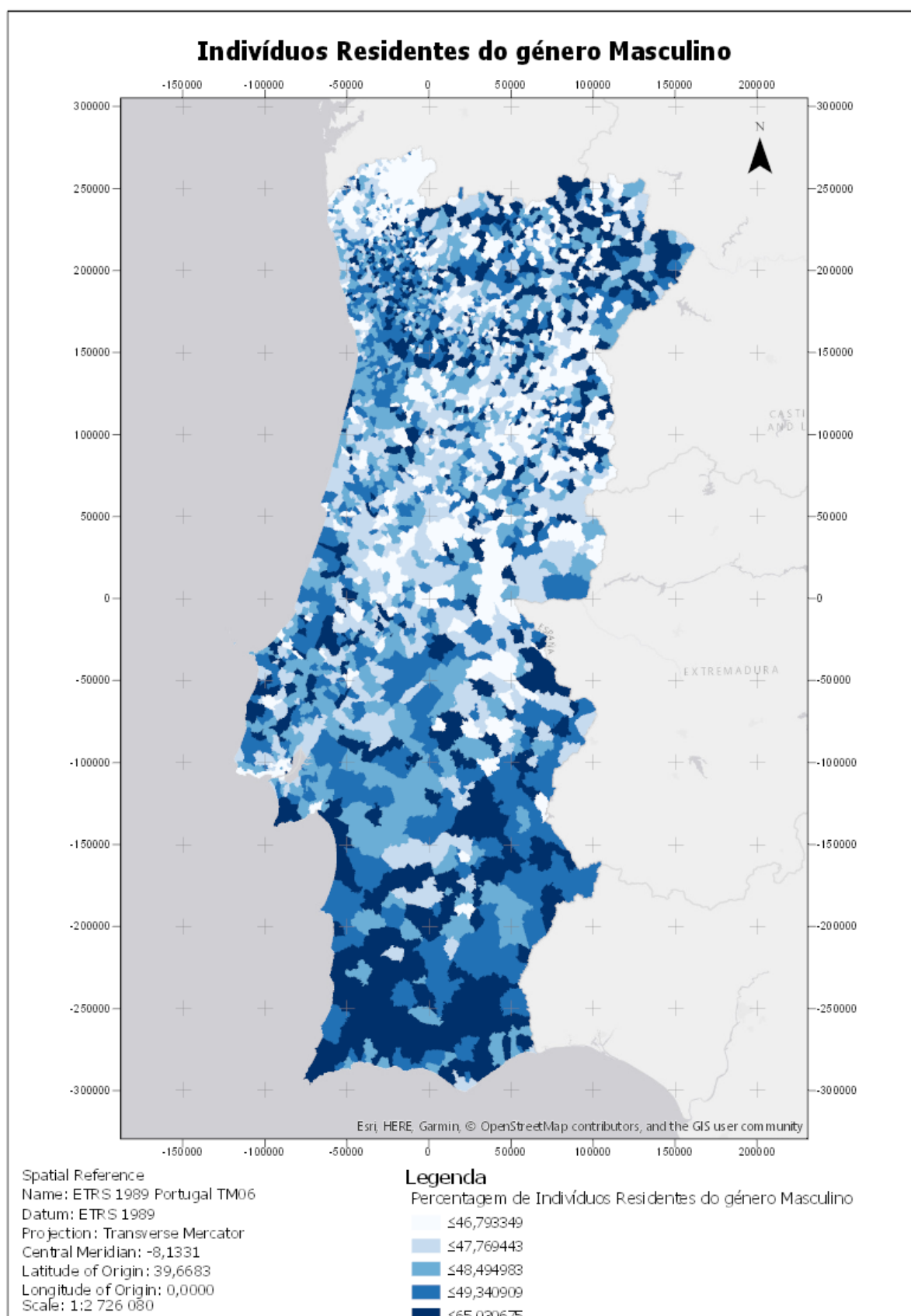
- relationships.htm#GUID-729B3B01-6911-41E9-AA99-8A4CF74EEE27. [Accessed: 24-Sep-2018].
- [37] D. Khan, L. M. Rossen, B. E. Hamilton, Y. He, R. Wei, and E. Dienes, “Hot spots, cluster detection and spatial outlier analysis of teen birth rates in the U.S., 2003–2012,” *Spat. Spatiotemporal. Epidemiol.*, vol. 21, pp. 67–75, 2017.
- [38] R. Huang, A. V. Moudon, A. J. Cook, and A. Drewnowski, “The spatial clustering of obesity: Does the built environment matter?,” *J. Hum. Nutr. Diet.*, vol. 28, no. 6, pp. 604–612, 2015.
- [39] ESRI, “Enrich Layer,” 2018. [Online]. Available: <http://pro.arcgis.com/en/pro-app/tool-reference/analysis/enrich-layer.htm>. [Accessed: 30-Aug-2018].
- [40] ESRI, “Esri Demographic Data Release Notes : Portugal,” no. June, pp. 1–3, 2018.
- [41] P. Santana, C. Costa, A. Loureiro, J. Raposo, and J. M. Boavida, “Geografias da Diabetes Mellitus em Portugal: Como as Condições do Contexto Influenciam o Risco de Morrer,” Porto, 2014.
- [42] Associação Protectora dos Diabetes de Portugal, “Complicações.” [Online]. Available: <https://www.apdp.pt/Diabetes/complicacoes#hipertensao-arterial>.
- [43] SNS, “Transparência --- Explore.” [Online]. Available: <https://transparencia.sns.gov.pt/explore/?sort=modified>.
- [44] SNS, “Entidades de Saúde - SNS,” 2017. [Online]. Available: <https://www.sns.gov.pt/institucional/entidades-de-saude/>. [Accessed: 20-Jul-2018].
- [45] Câmara Municipal de Lisboa, “Geodados.” [Online]. Available: <http://geodados.cm-lisboa.pt/>.
- [46] MB-Research, “Methodological Notes and Usage of MB-Research Market Data,” p. 1, 2016.
- [47] Direcção Geral do Território (DGT), “DGT - CAOP,” 2017. [Online]. Available: [http://www.dgterritorio.pt/cartografia\\_e\\_geodesia/cartografia/carta\\_administrativa\\_oficial\\_de\\_portugal\\_caop\\_/caop\\_\\_download\\_/](http://www.dgterritorio.pt/cartografia_e_geodesia/cartografia/carta_administrativa_oficial_de_portugal_caop_/caop__download_/). [Accessed: 20-Jul-2018].
- [48] ESRI, “Service area analysis - Help |ArcGIS Desktop.” [Online]. Available: [http://desktop.arcgis.com/en/arcmap/latest/extensions/network-analyst/service-area.htm#ESRI\\_SECTION1\\_DEC185407C154056B6B17077F9BAD8F2](http://desktop.arcgis.com/en/arcmap/latest/extensions/network-analyst/service-area.htm#ESRI_SECTION1_DEC185407C154056B6B17077F9BAD8F2). [Accessed: 25-Sep-2018].
- [49] Sistema Nacional de Saúde (SNS), “GeoSaúde.” [Online]. Available: <http://www.geosaude.dgs.pt/websig/v5/portal2/public/index.php?par=geosaude>.

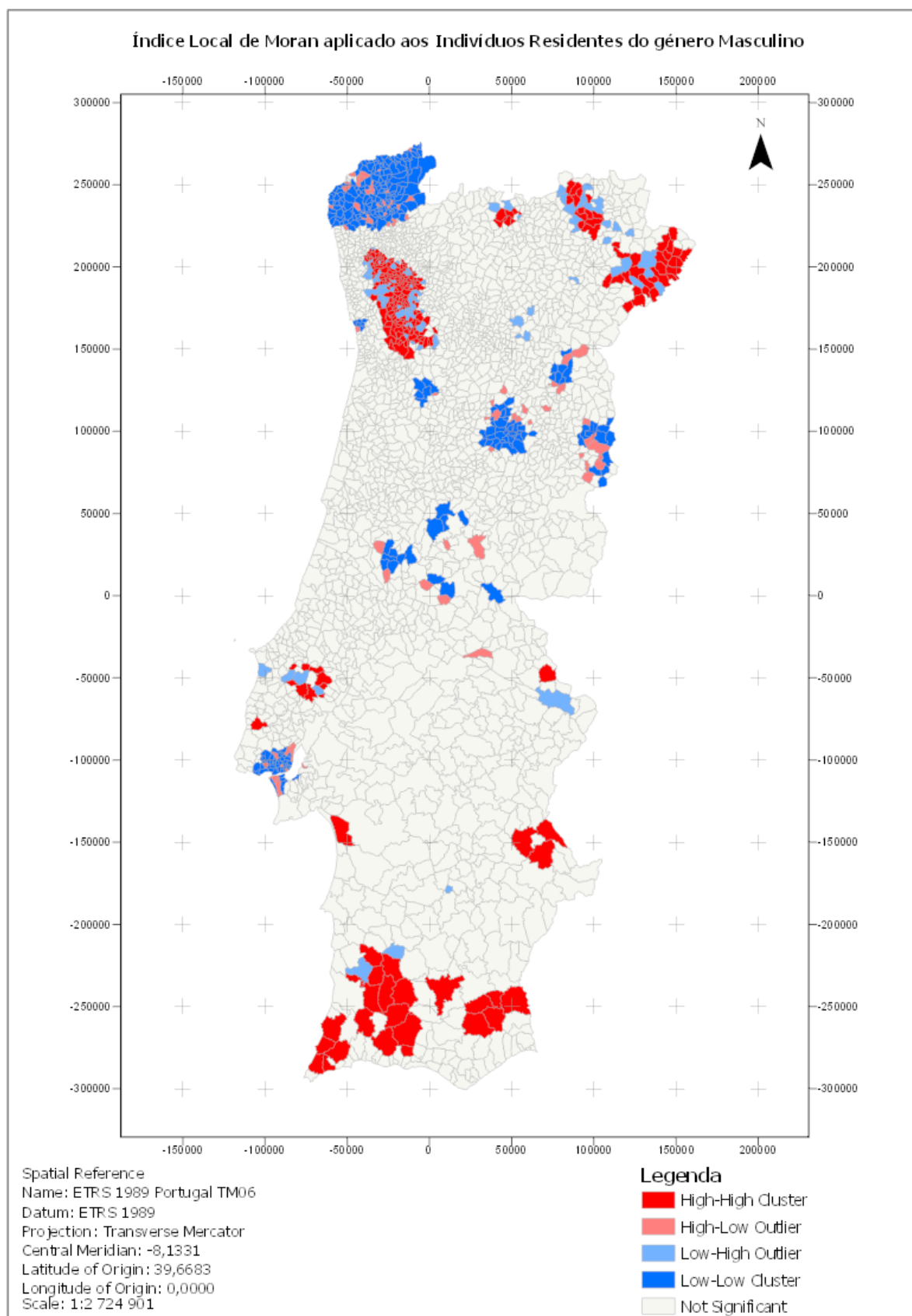
## Anexos

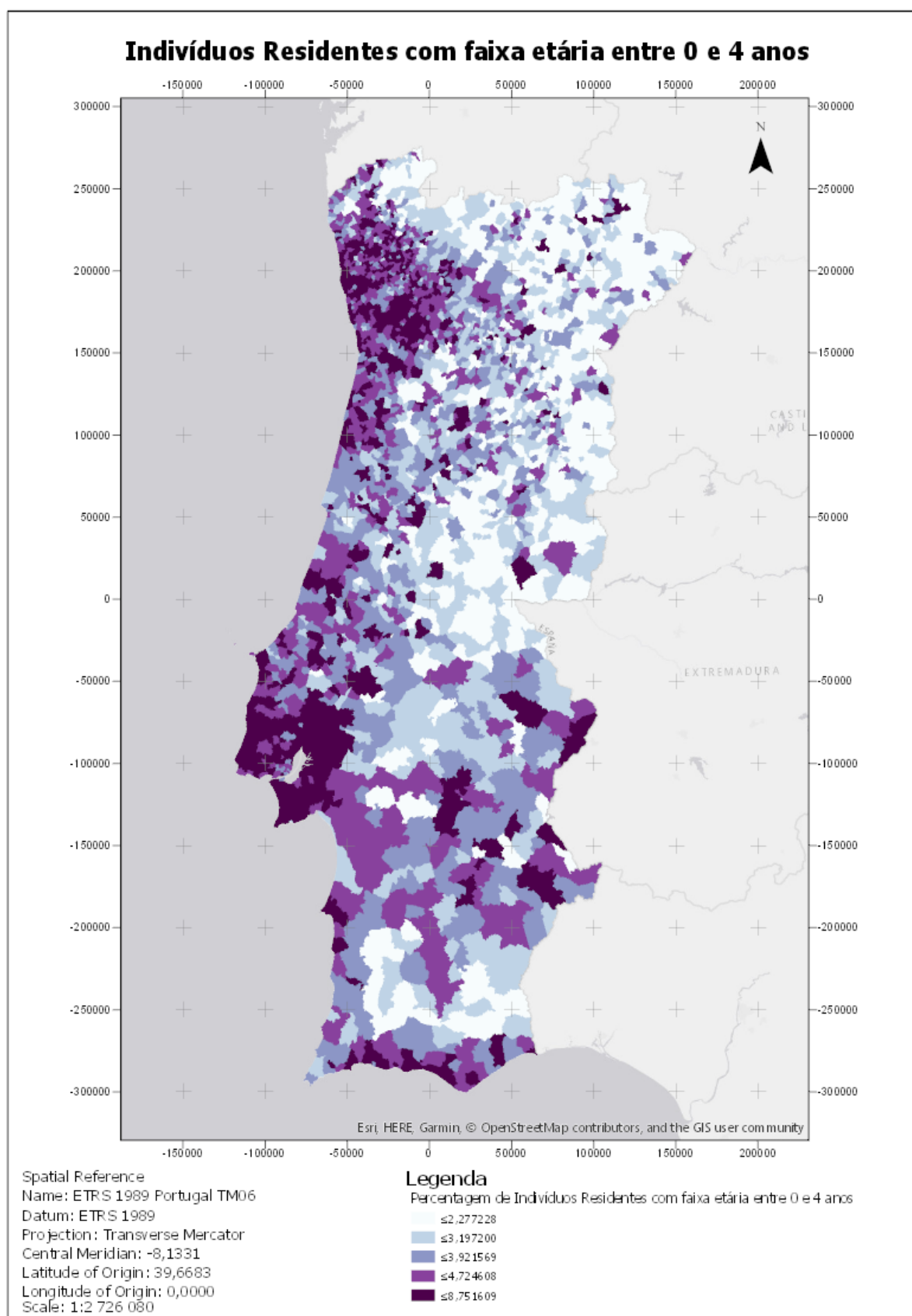
### Anexo I – Mapas com a Distribuição Espacial dos Atributos Utilizados na Segmentação



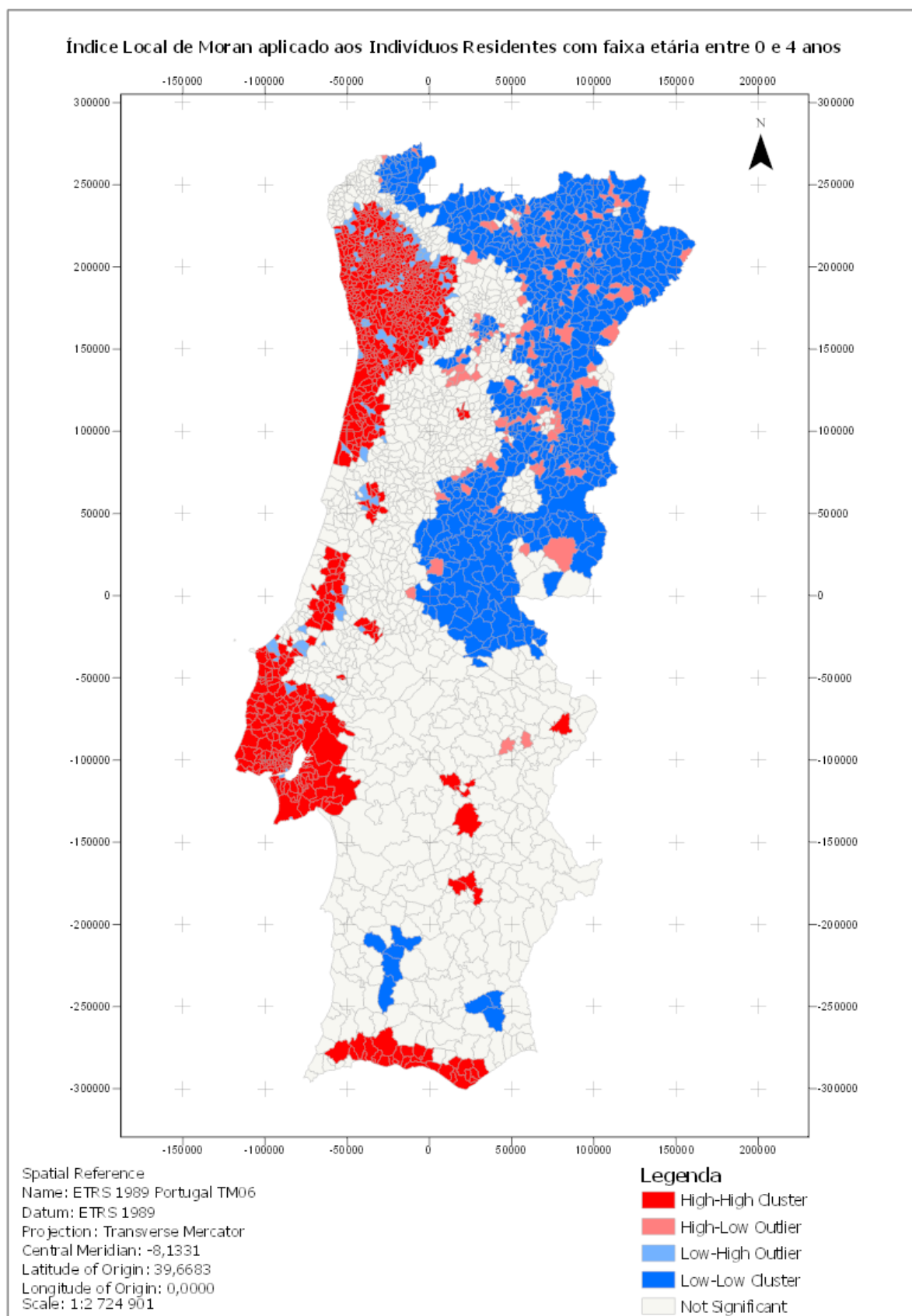


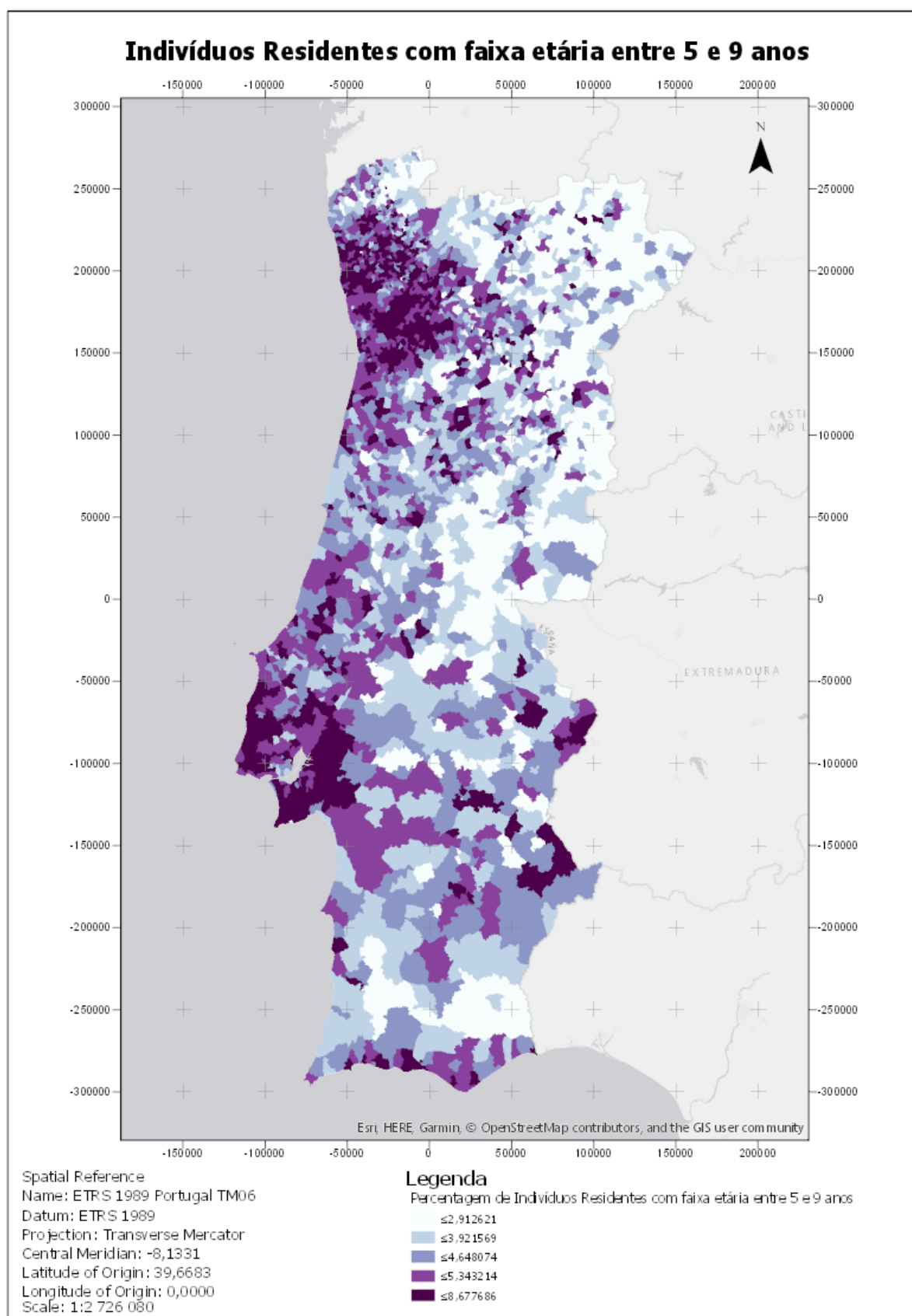


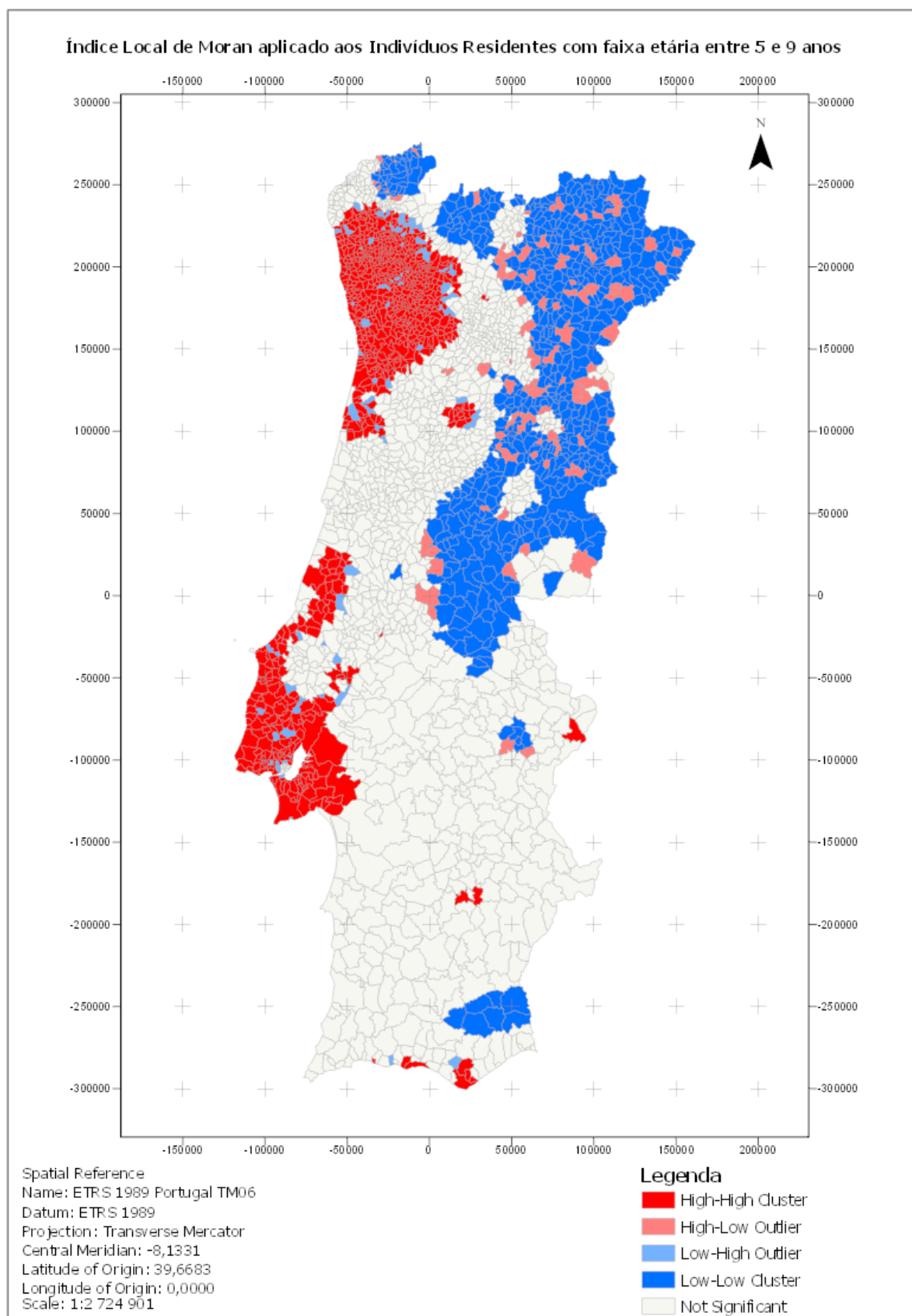


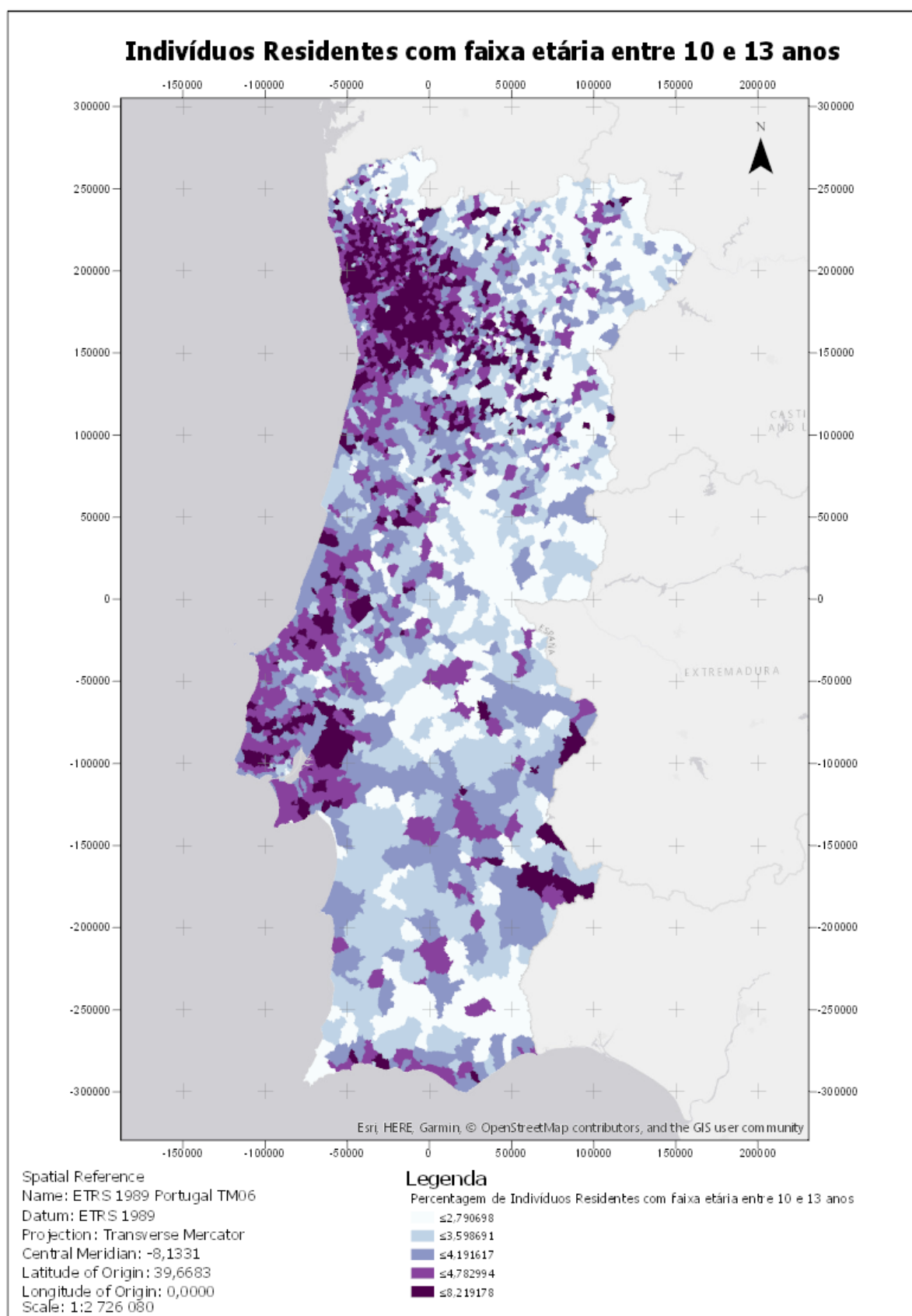


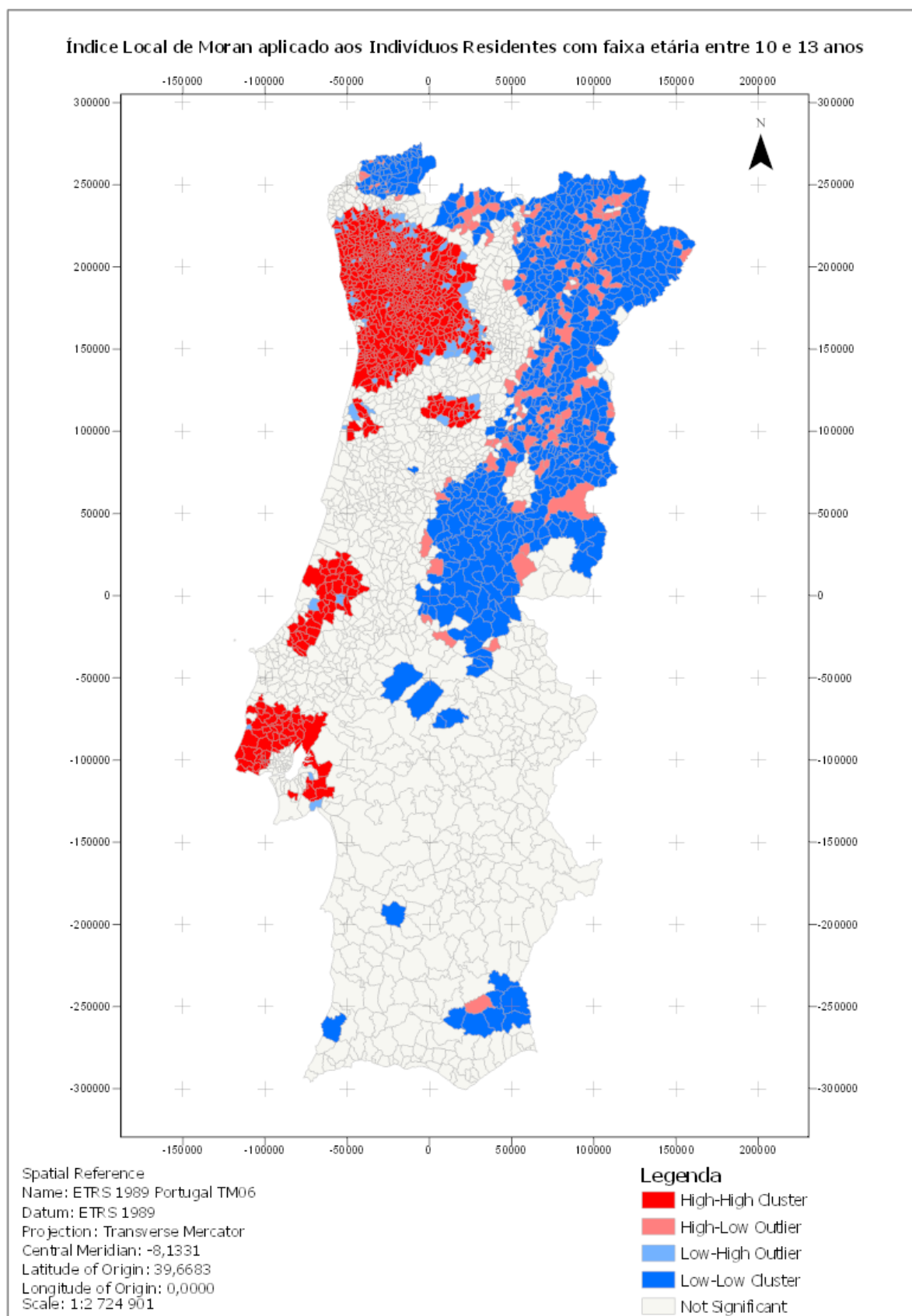


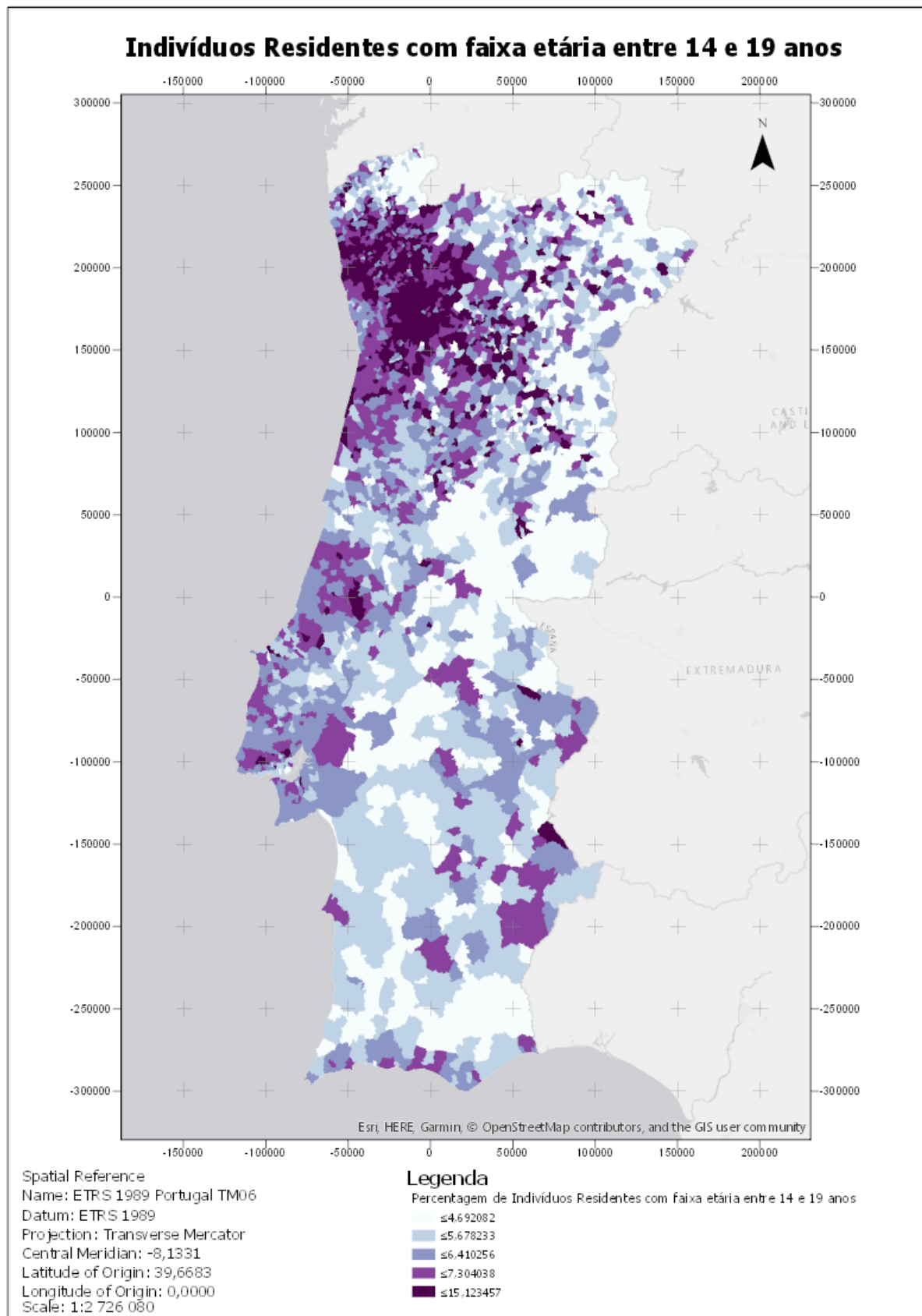




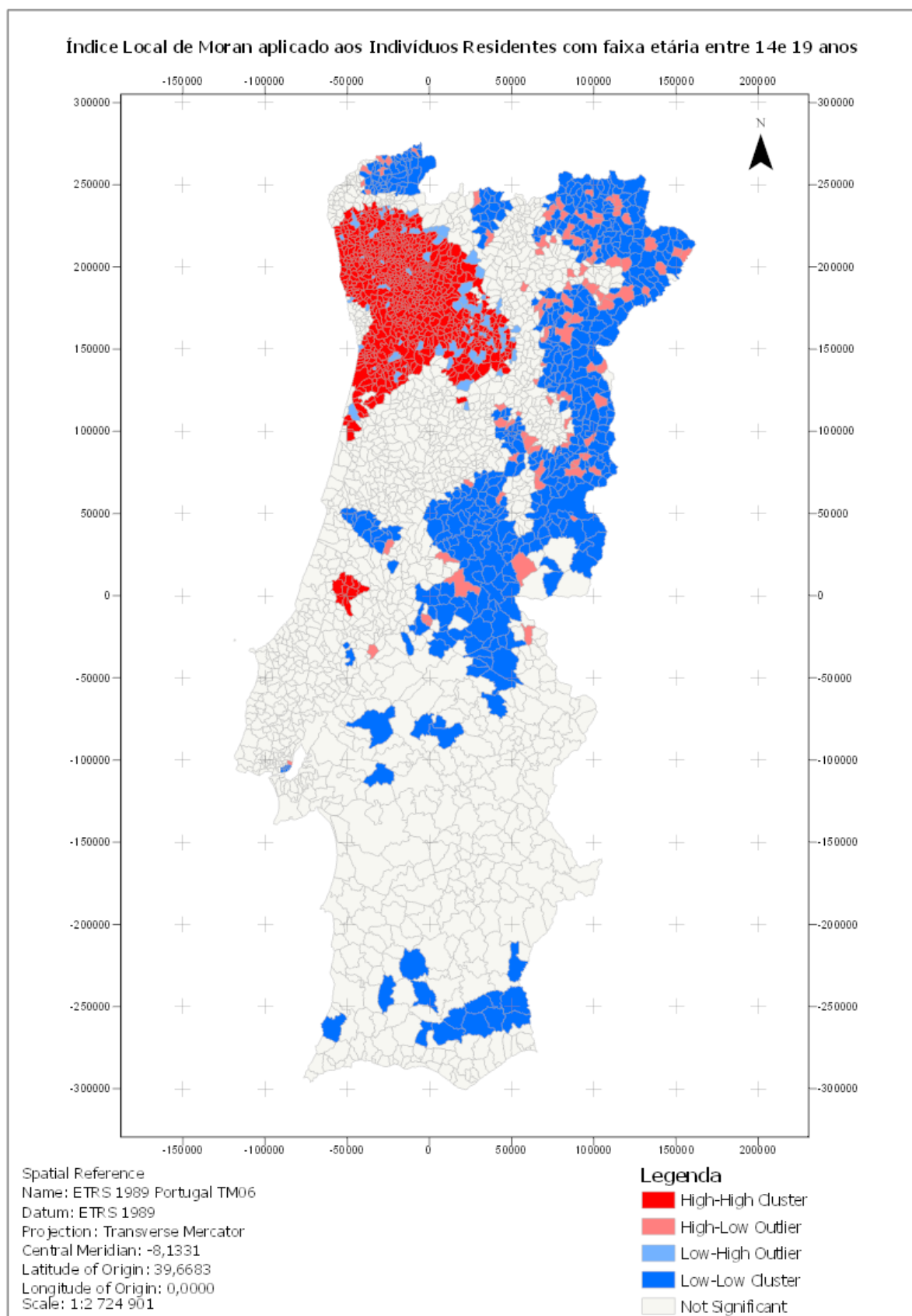




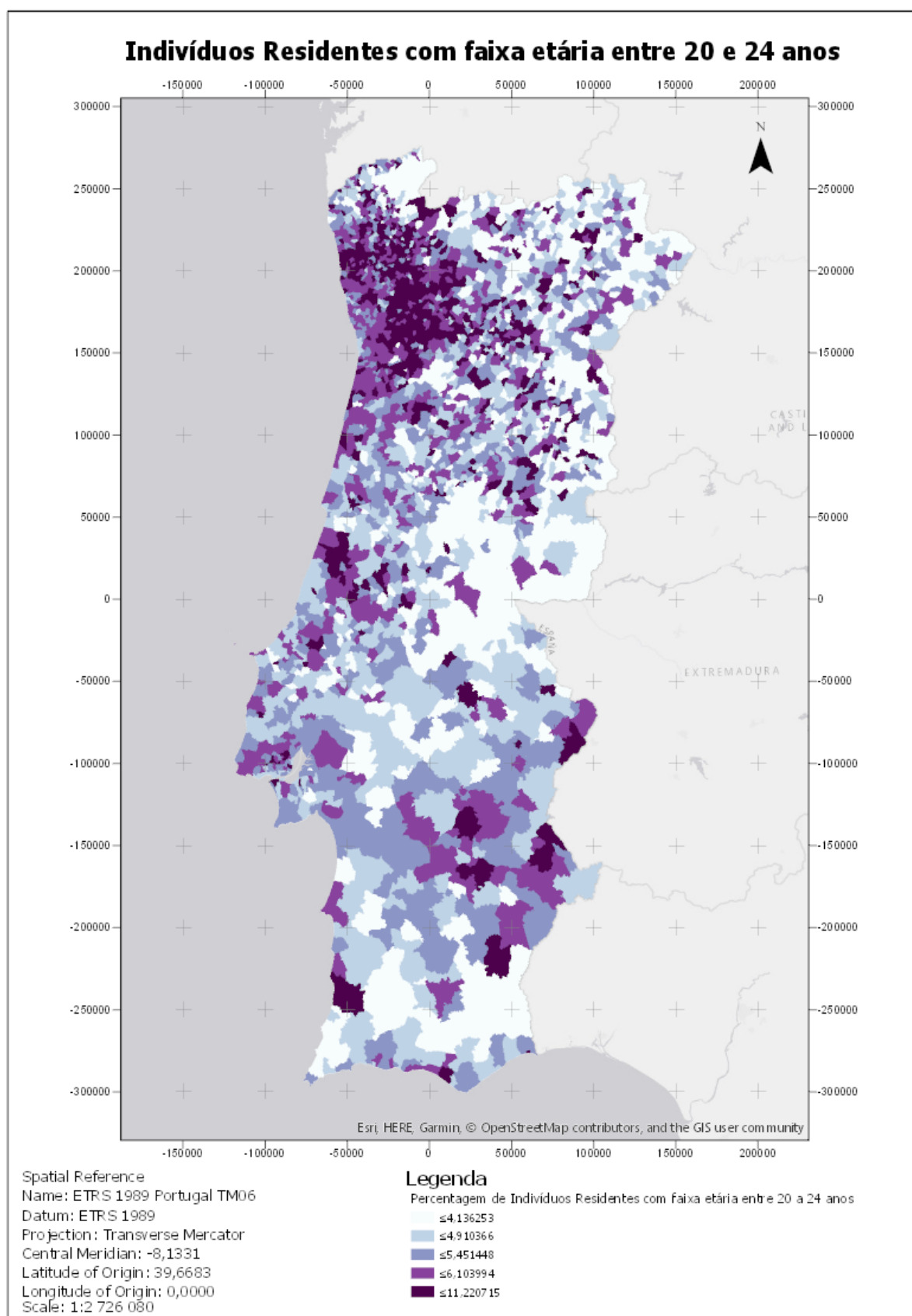


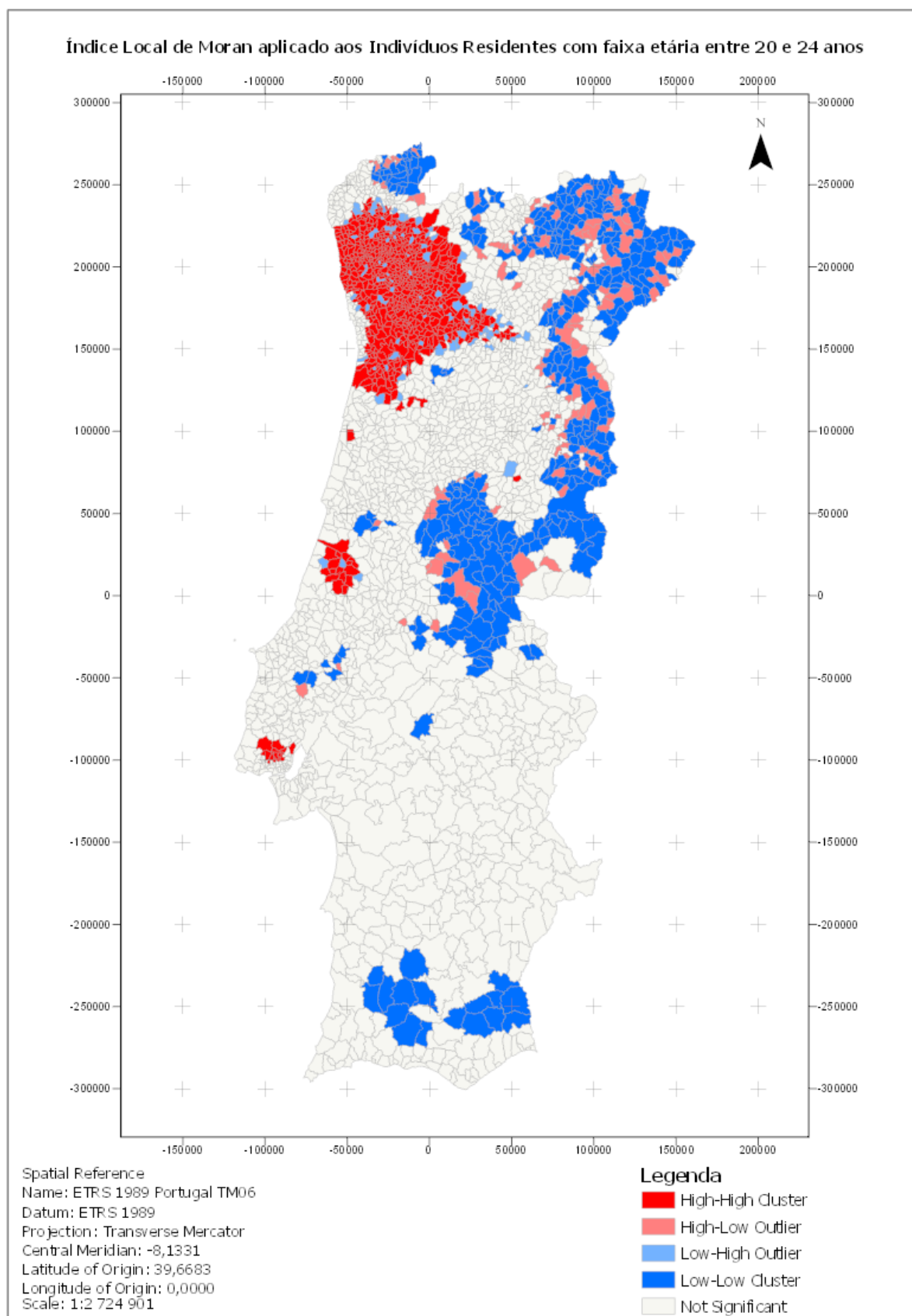


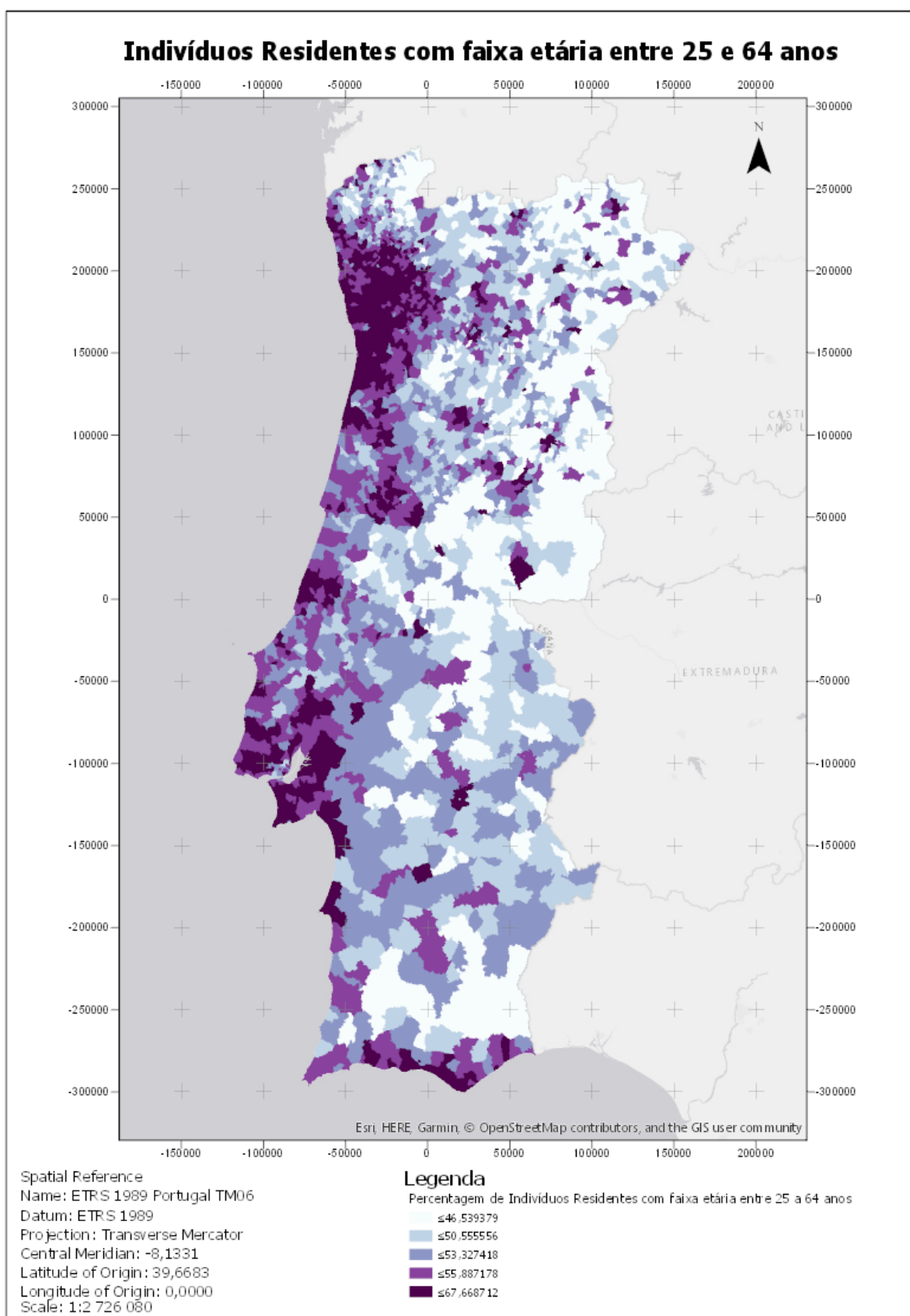


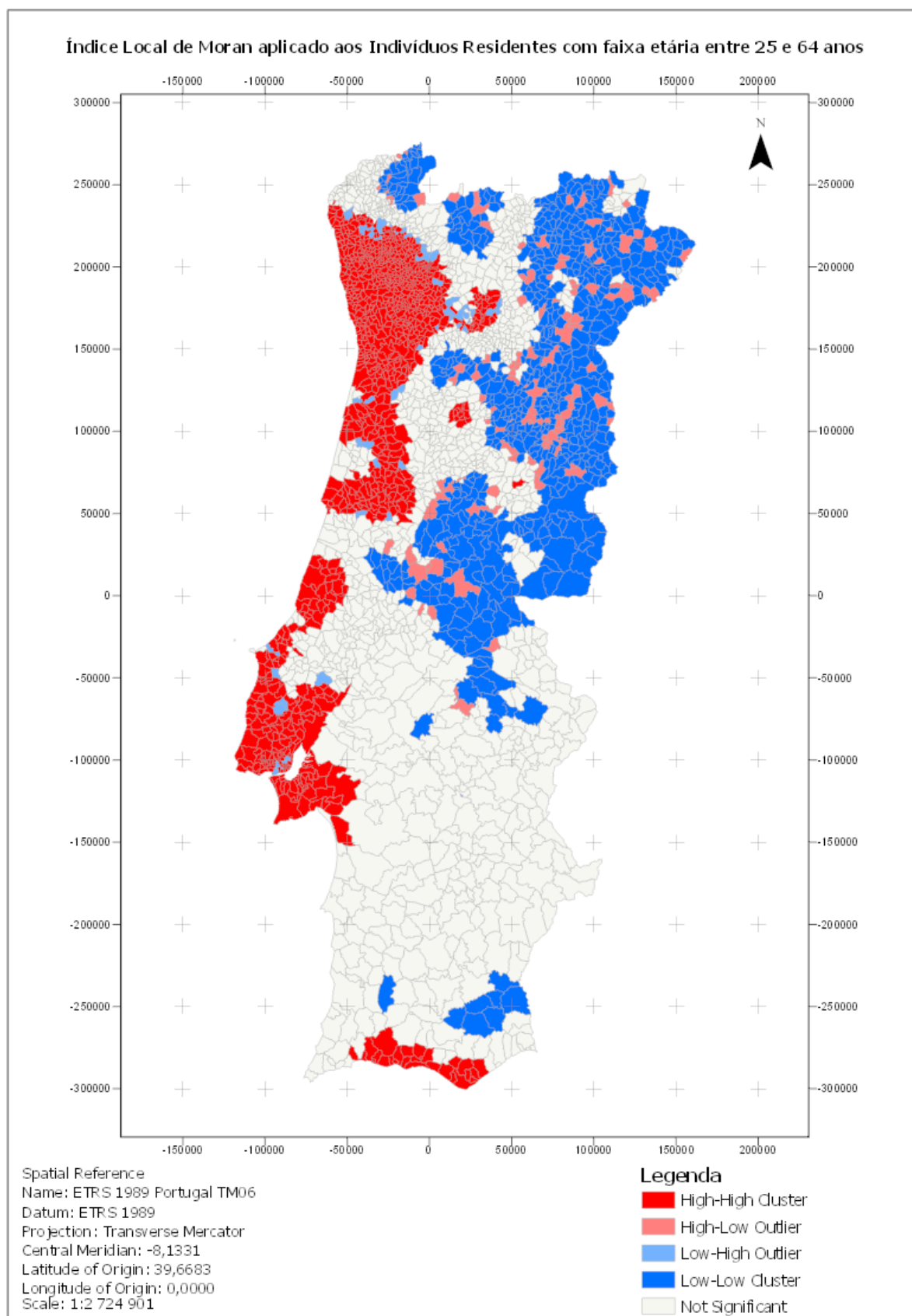


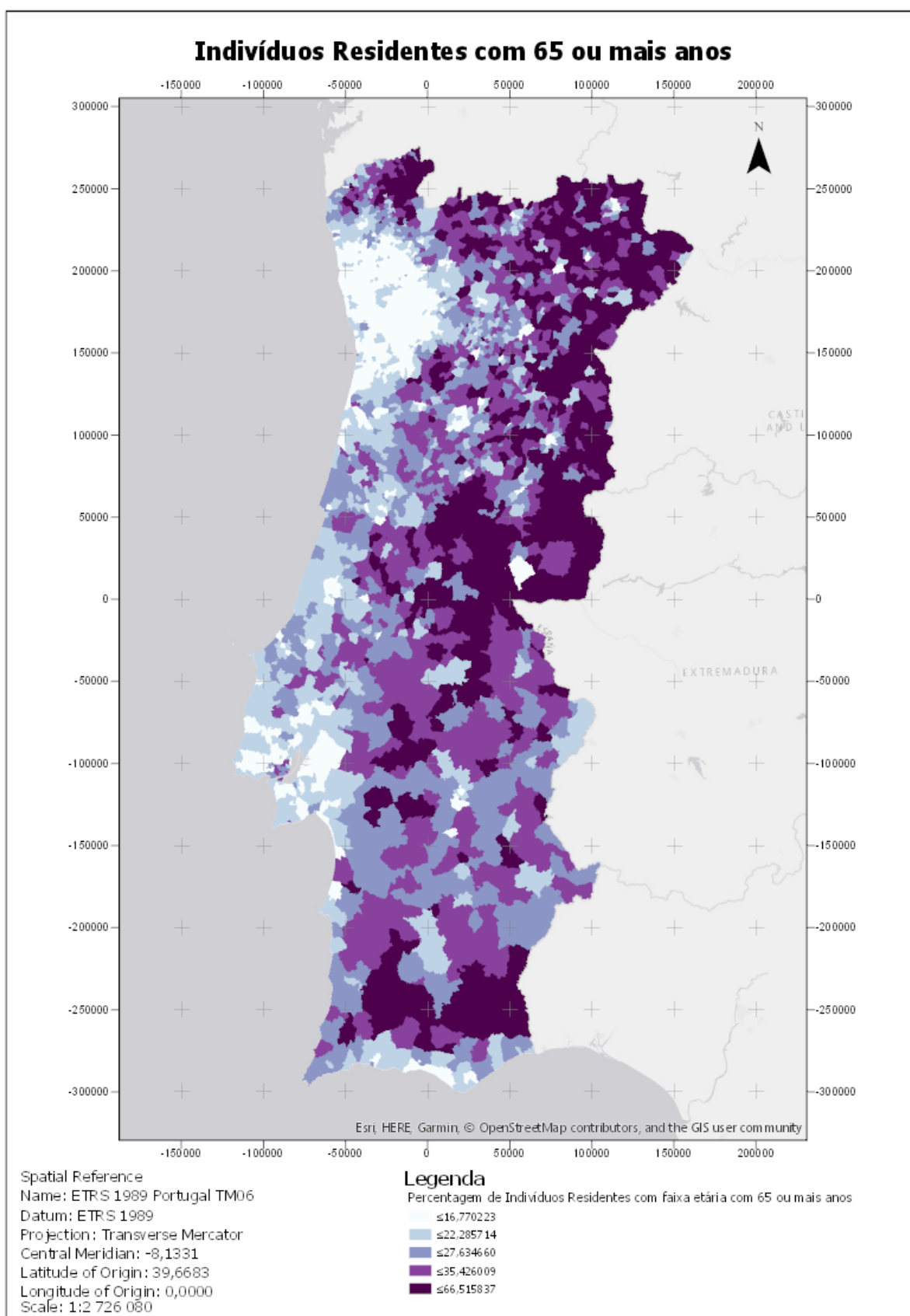


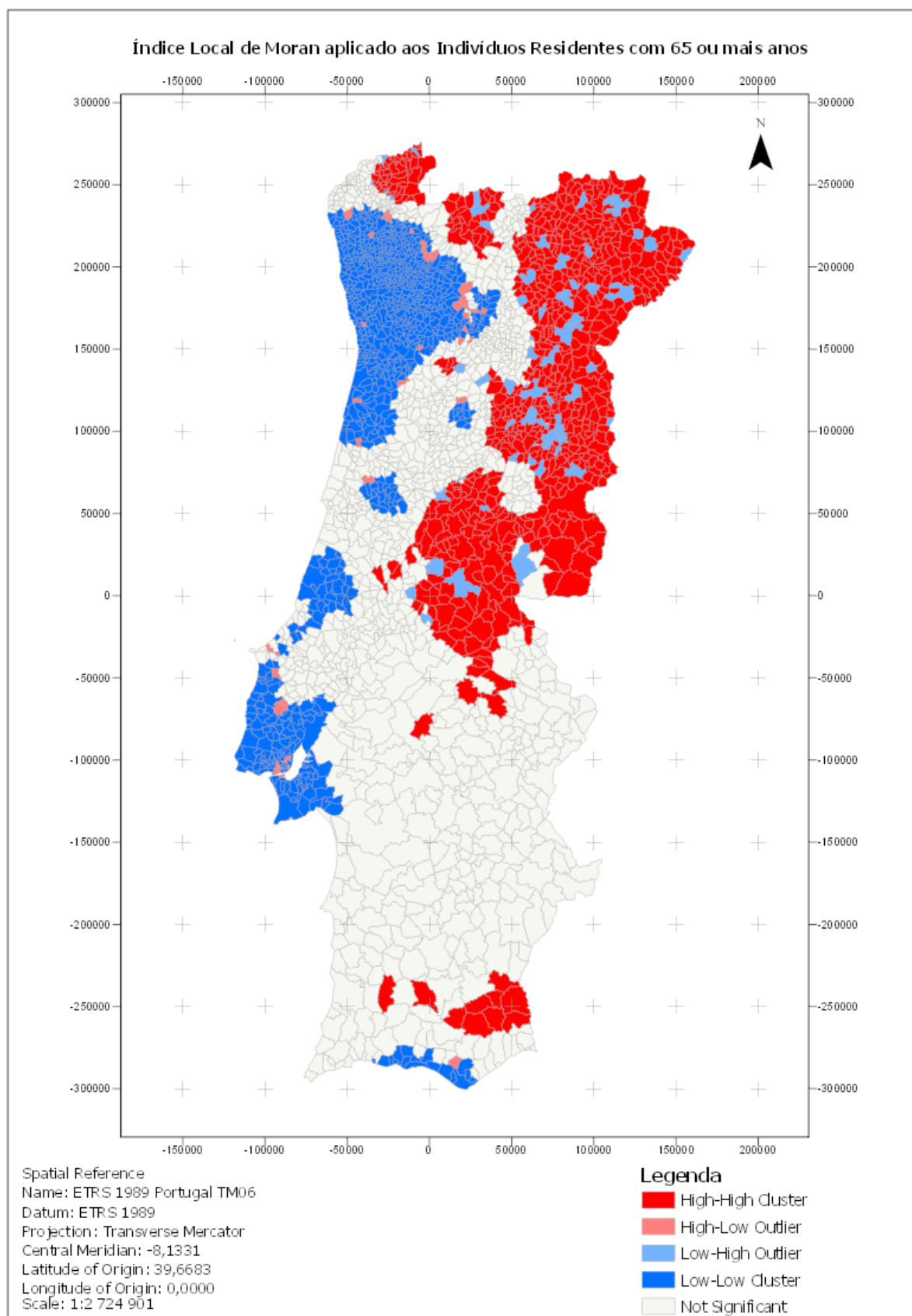




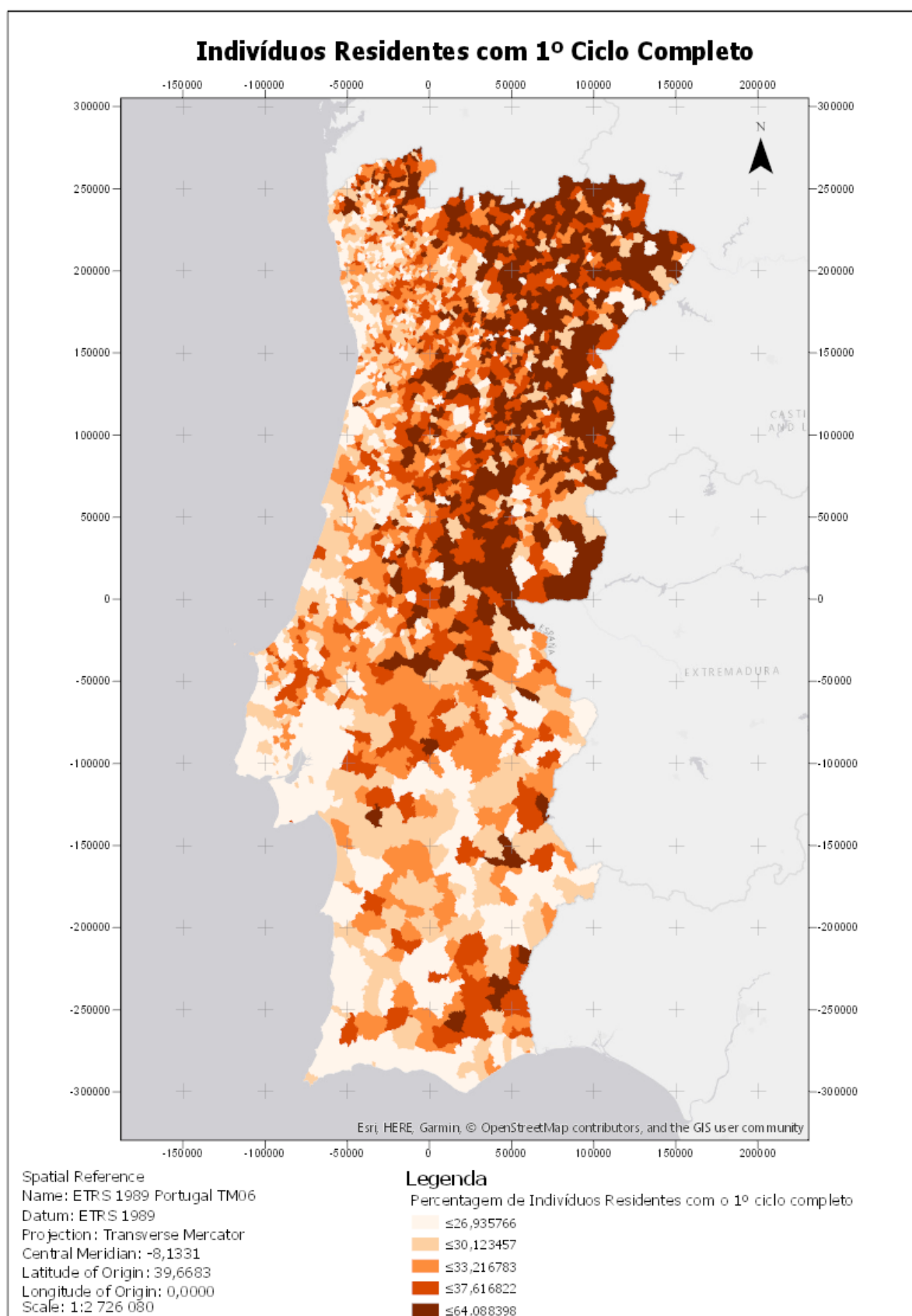




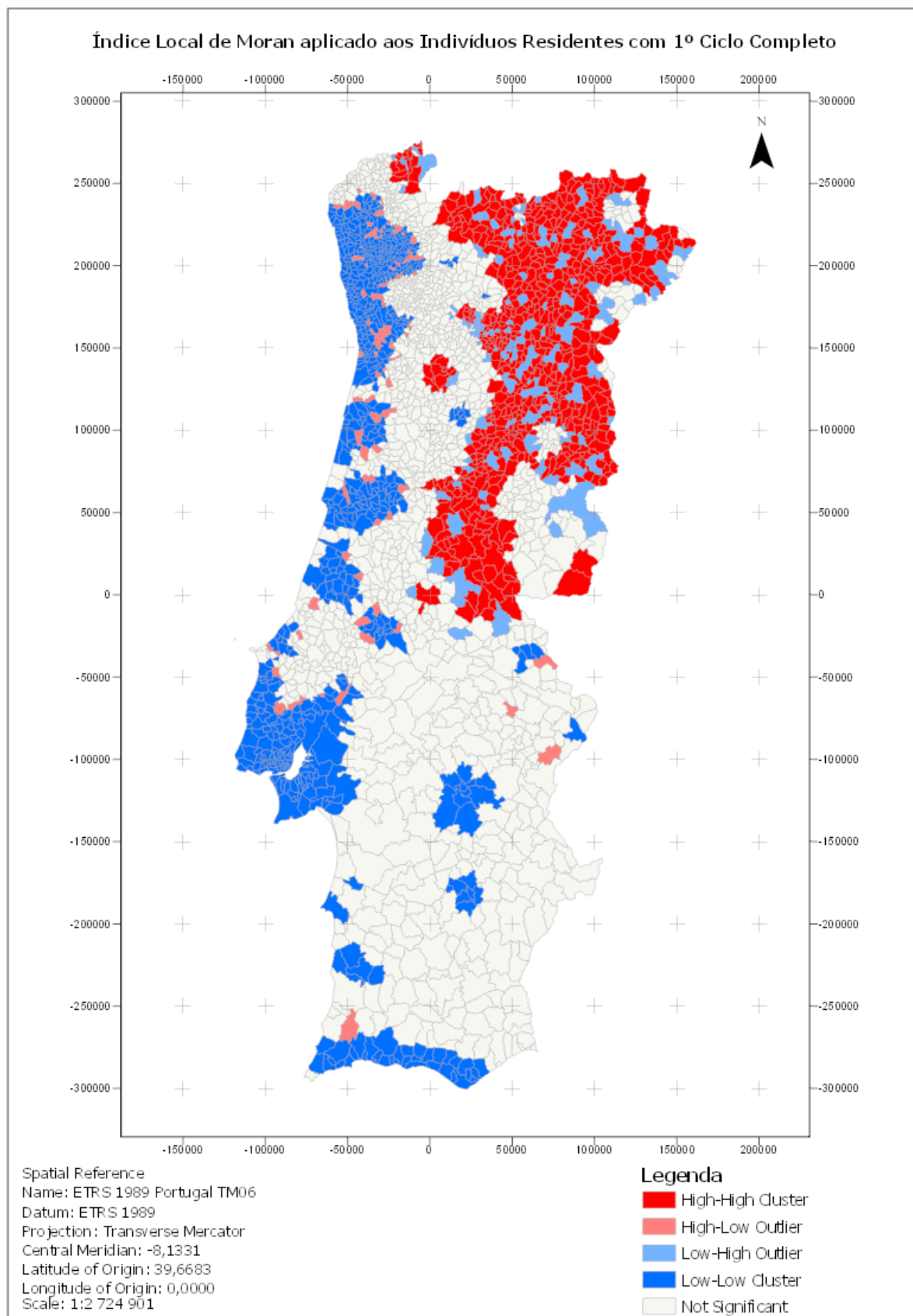


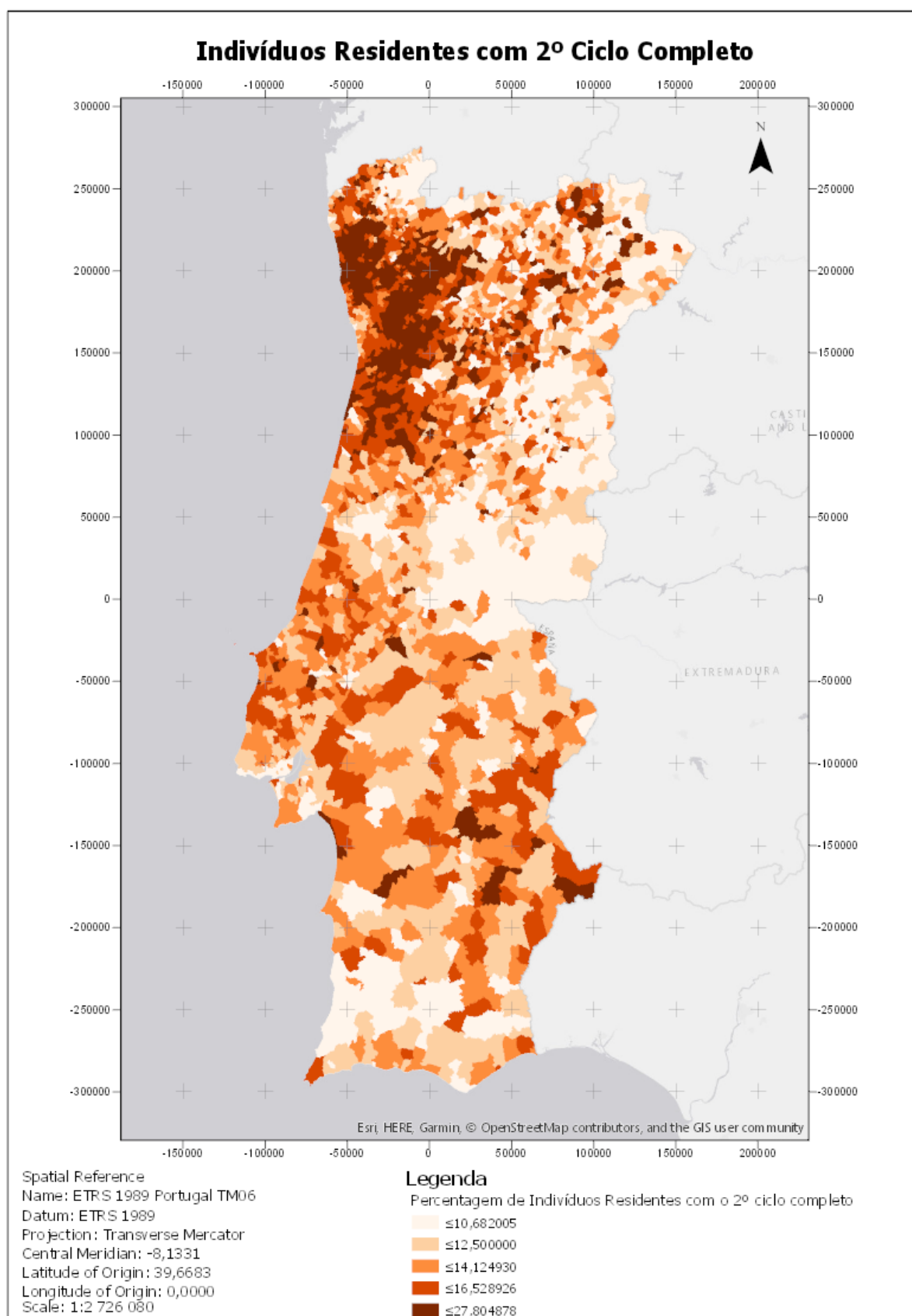


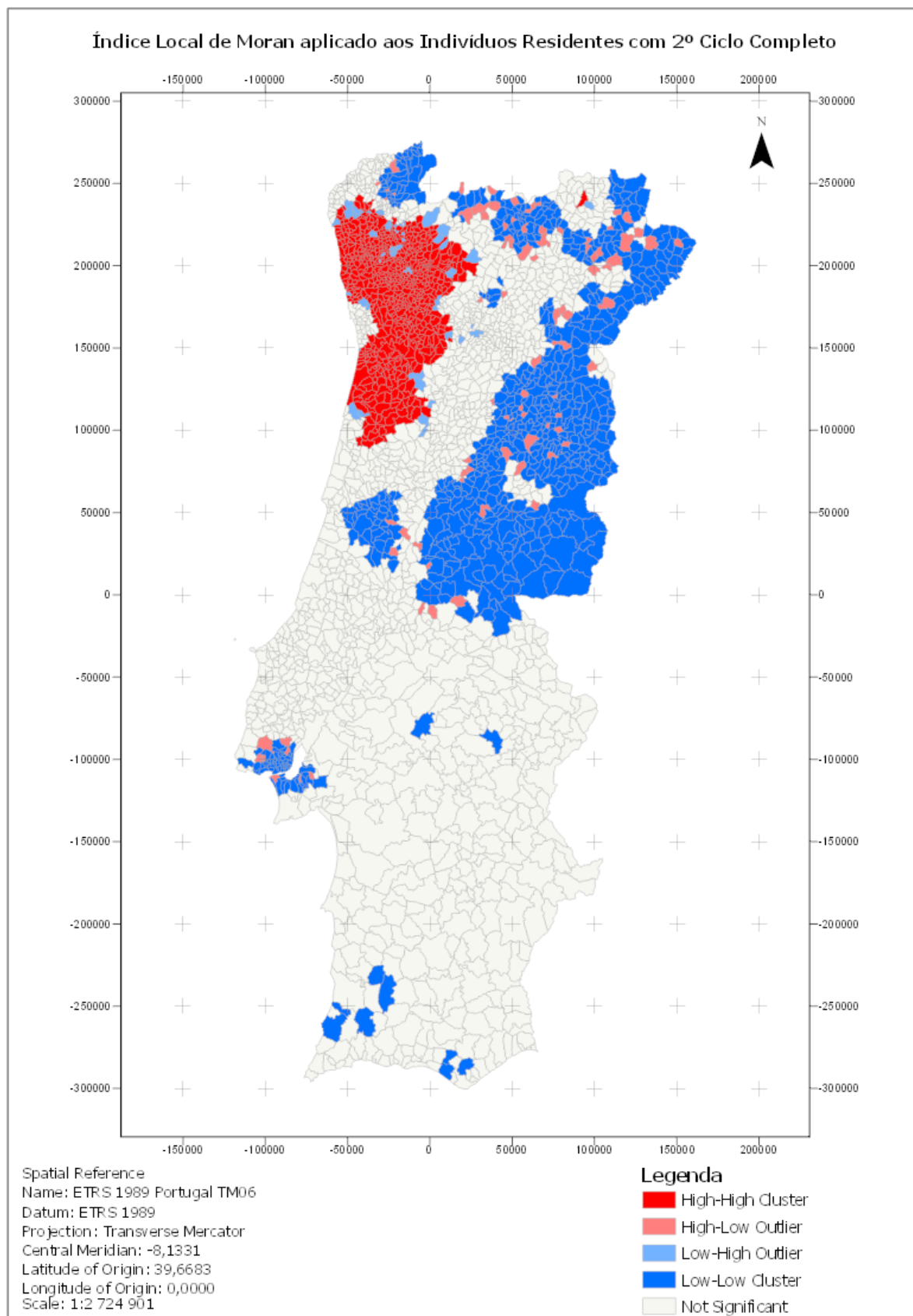


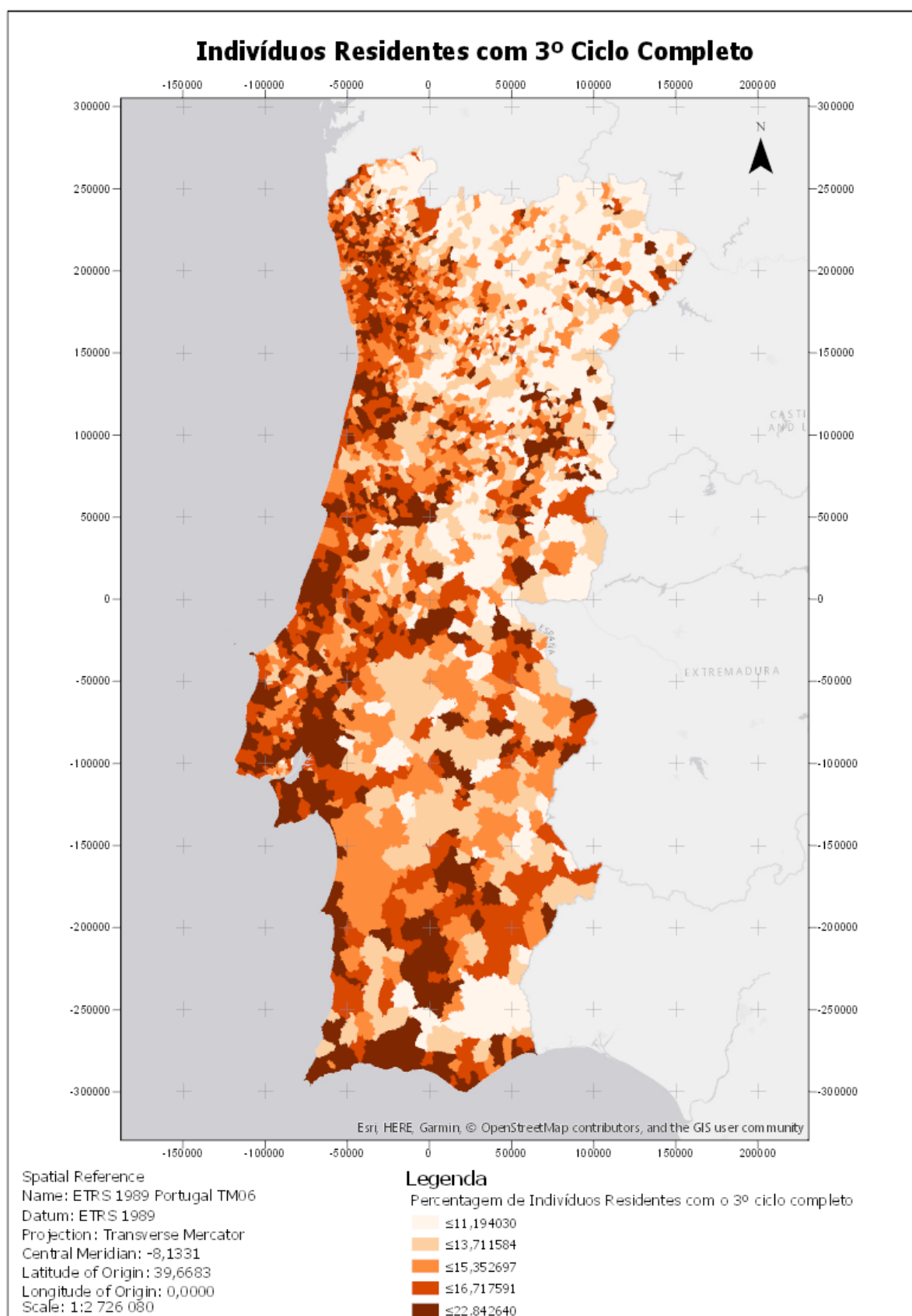


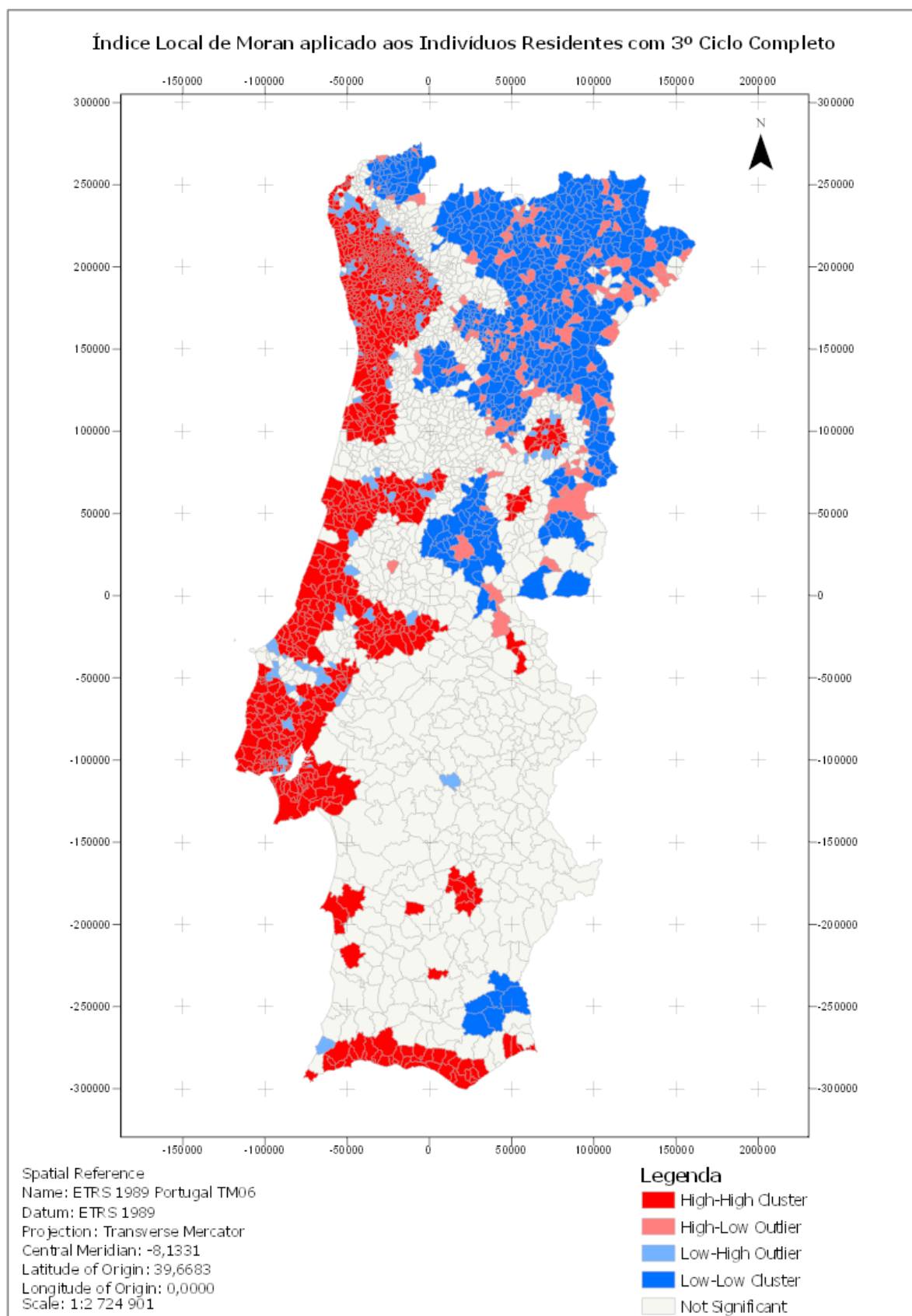


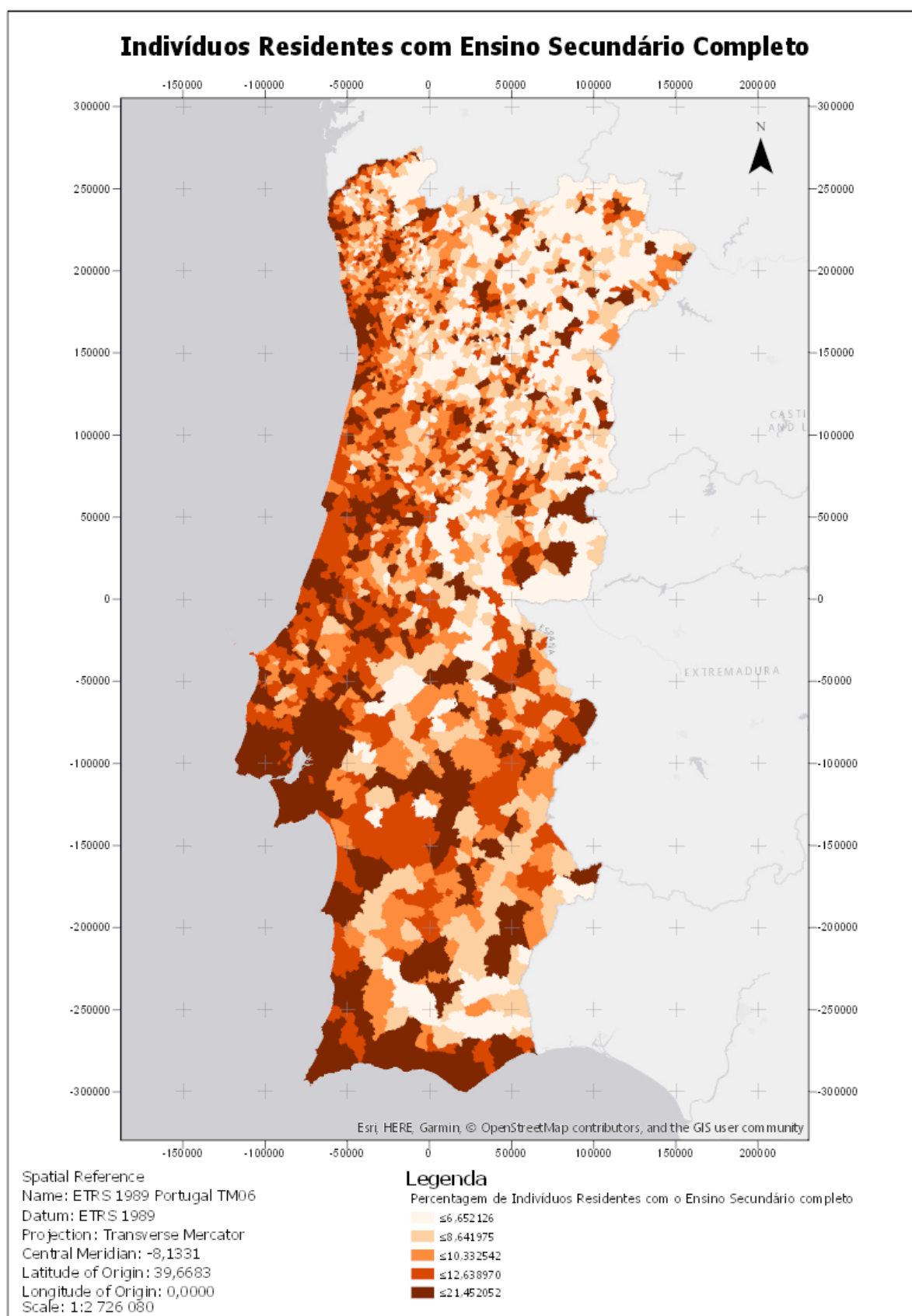




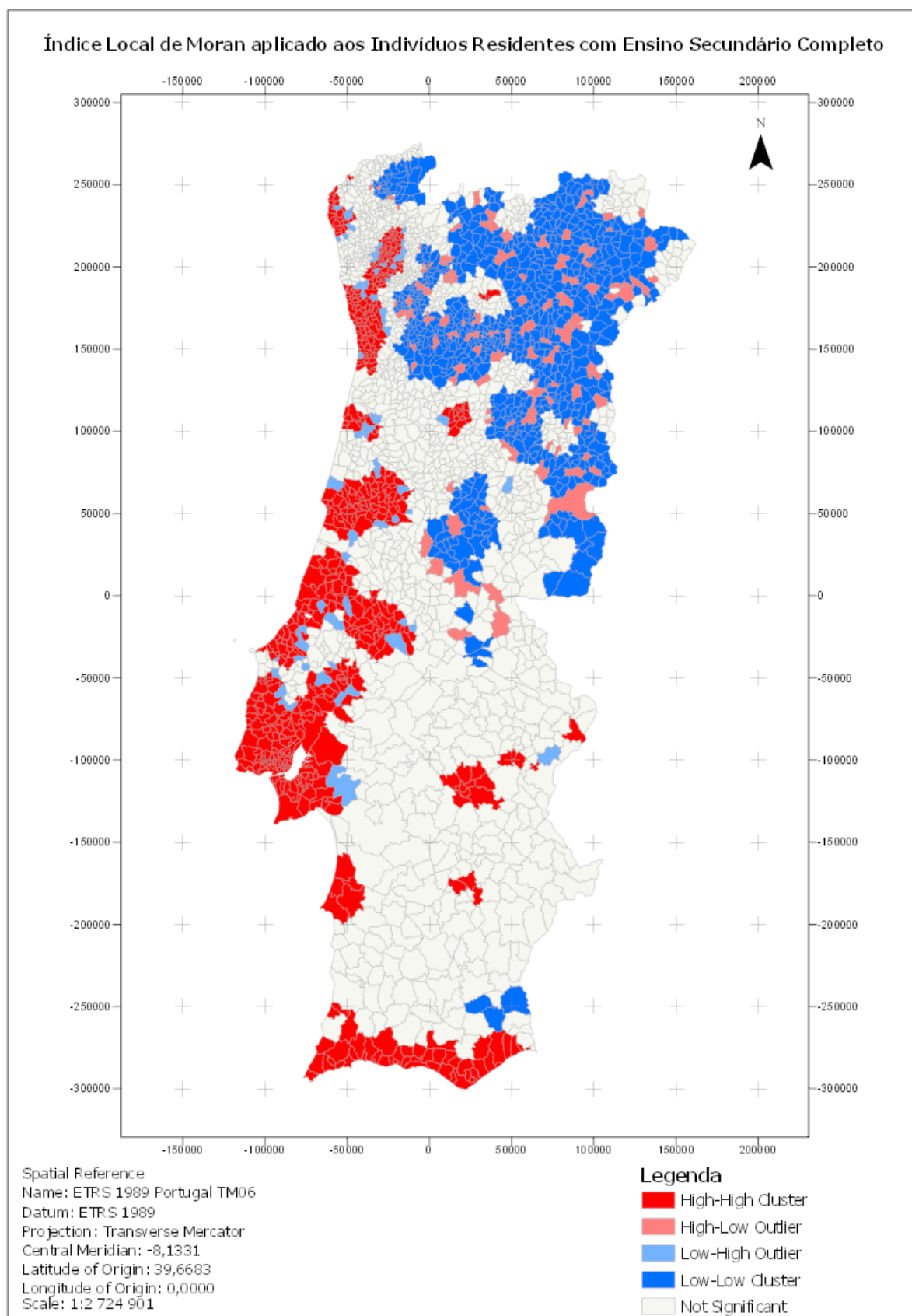




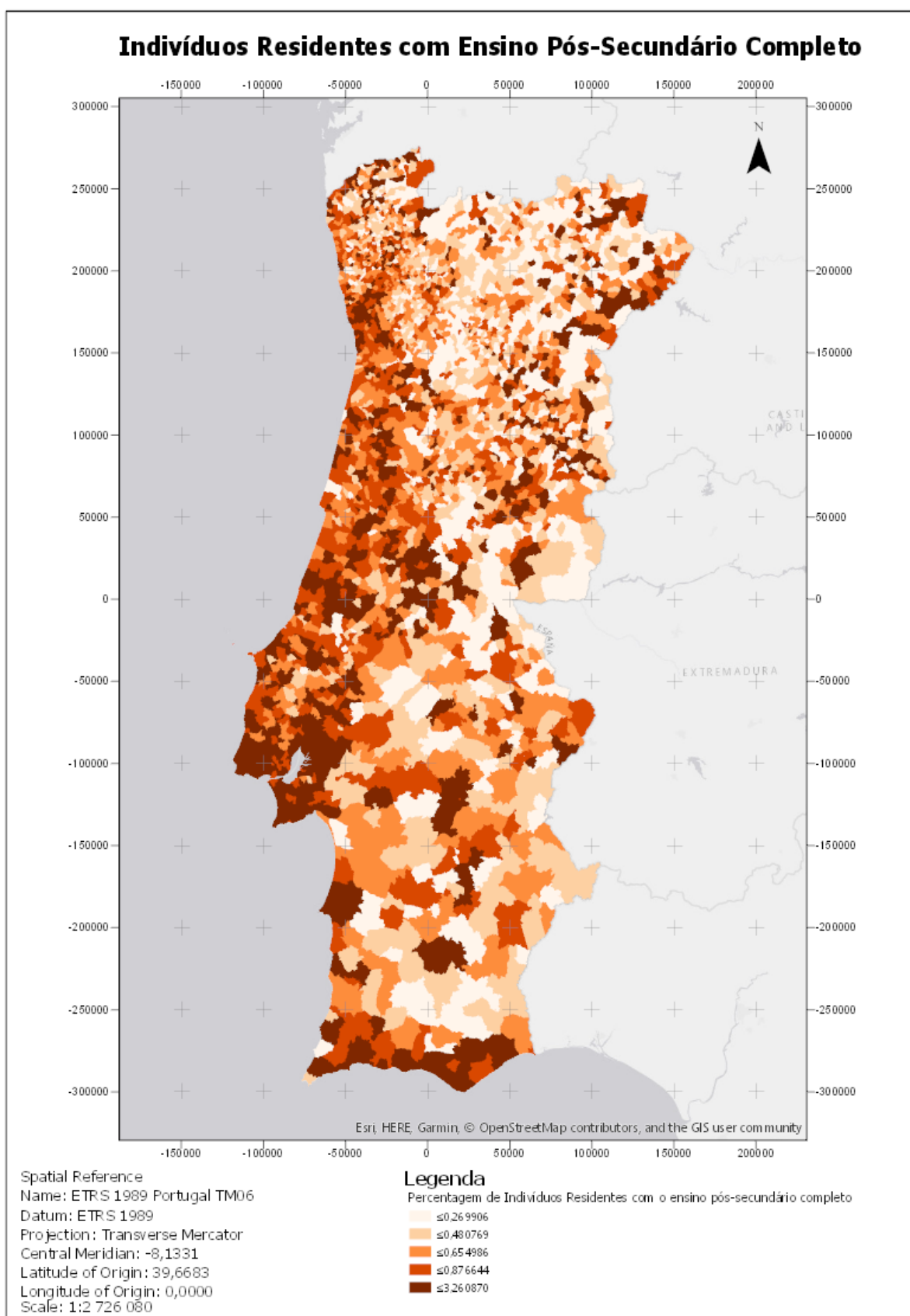


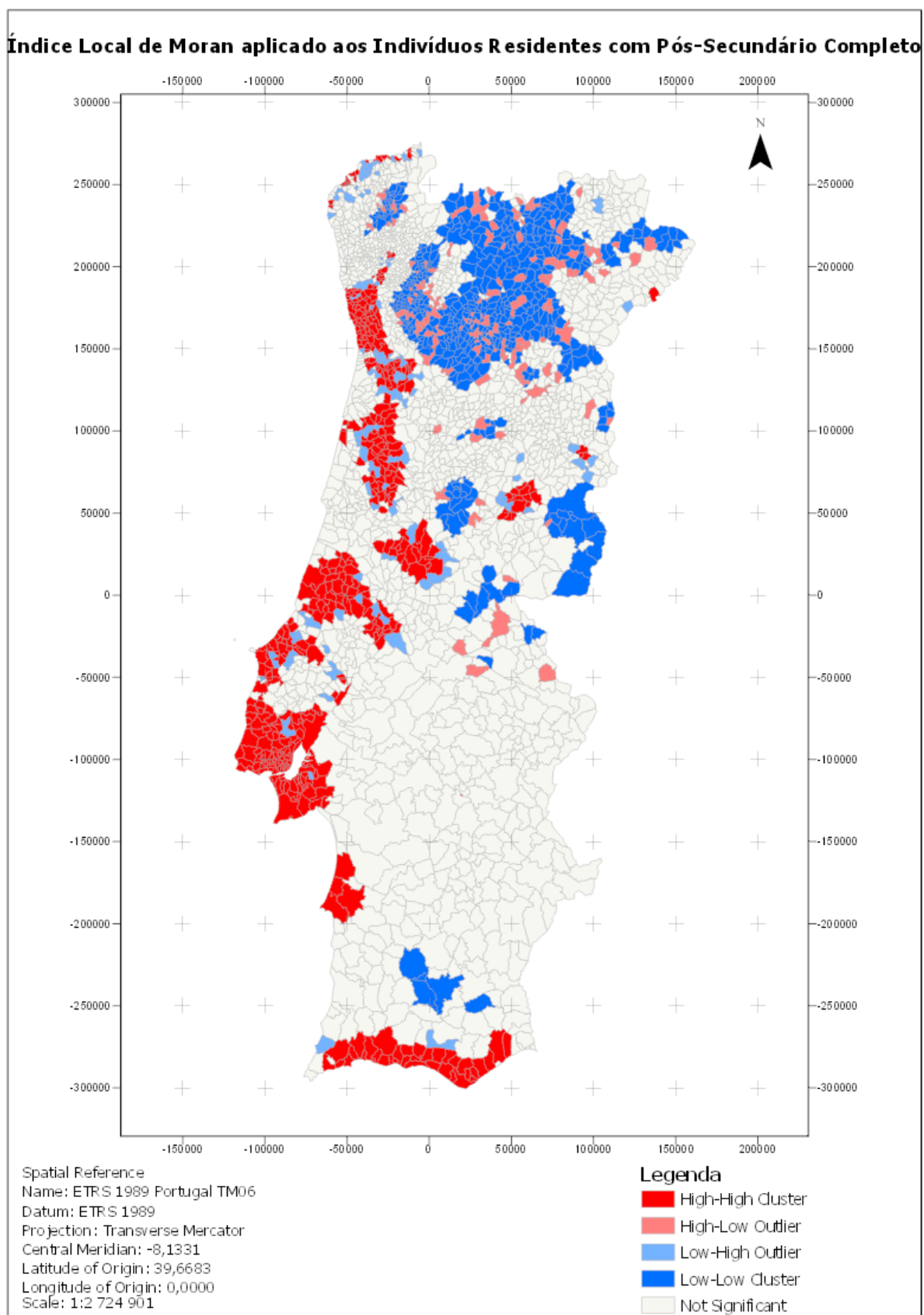


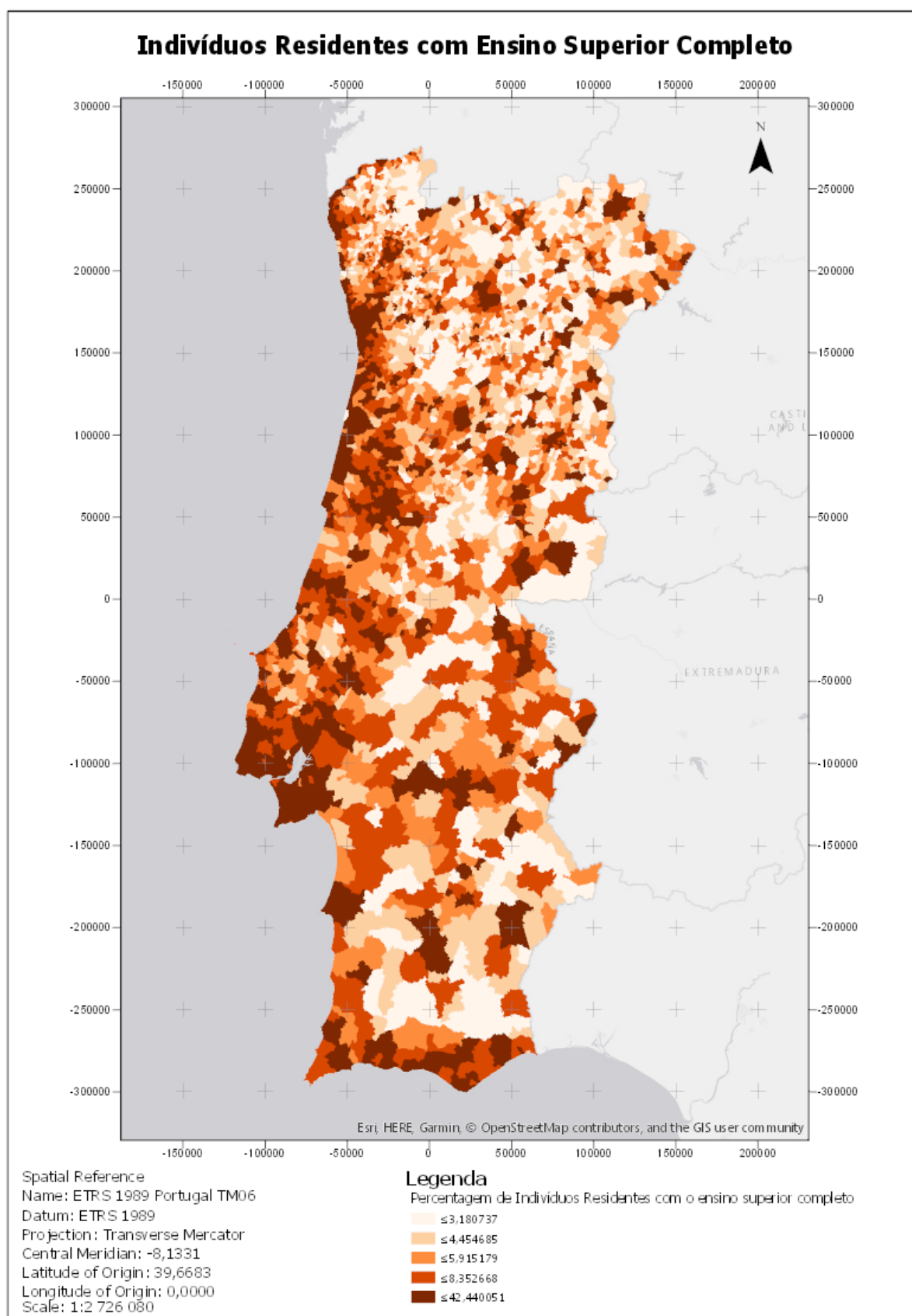


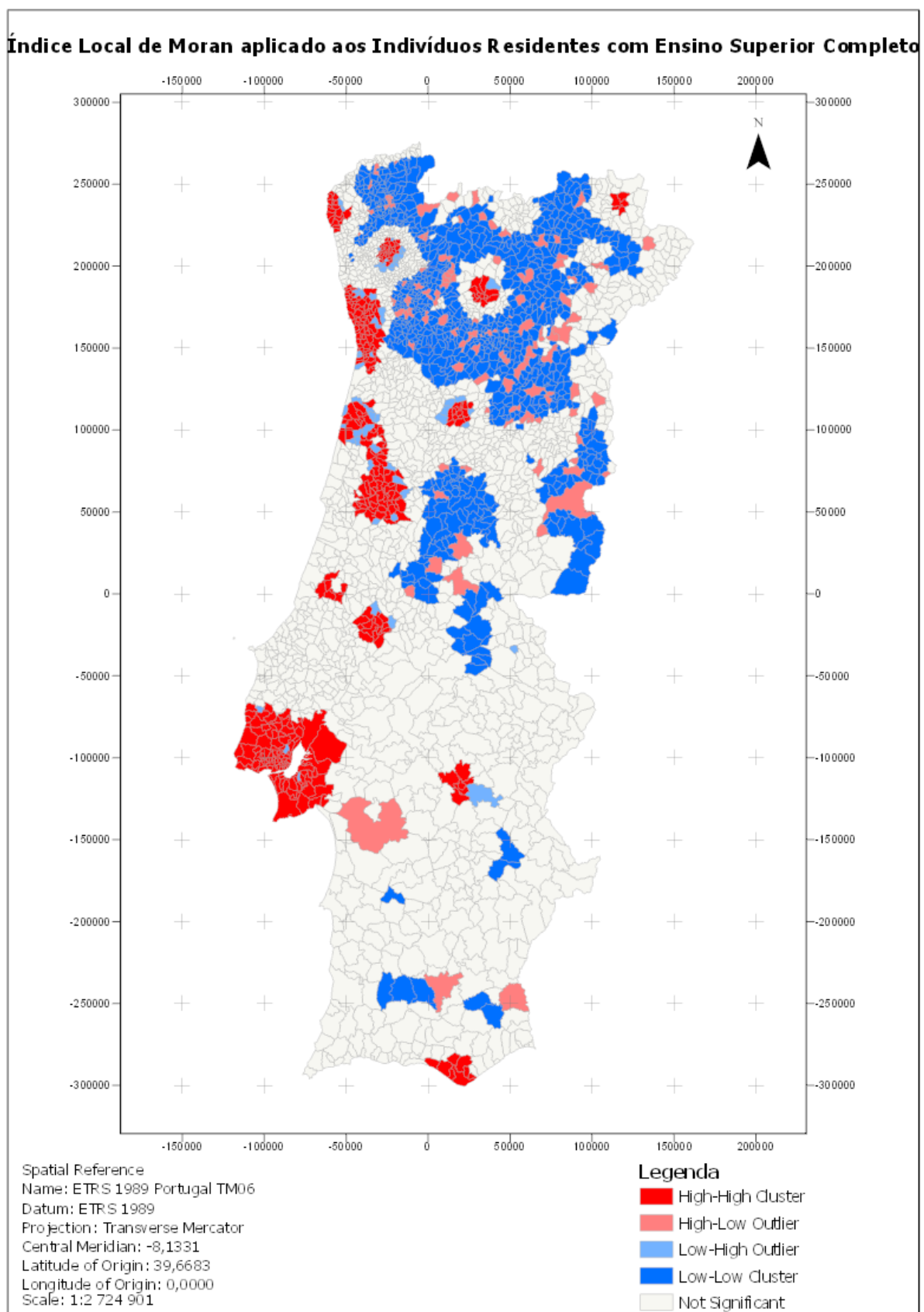


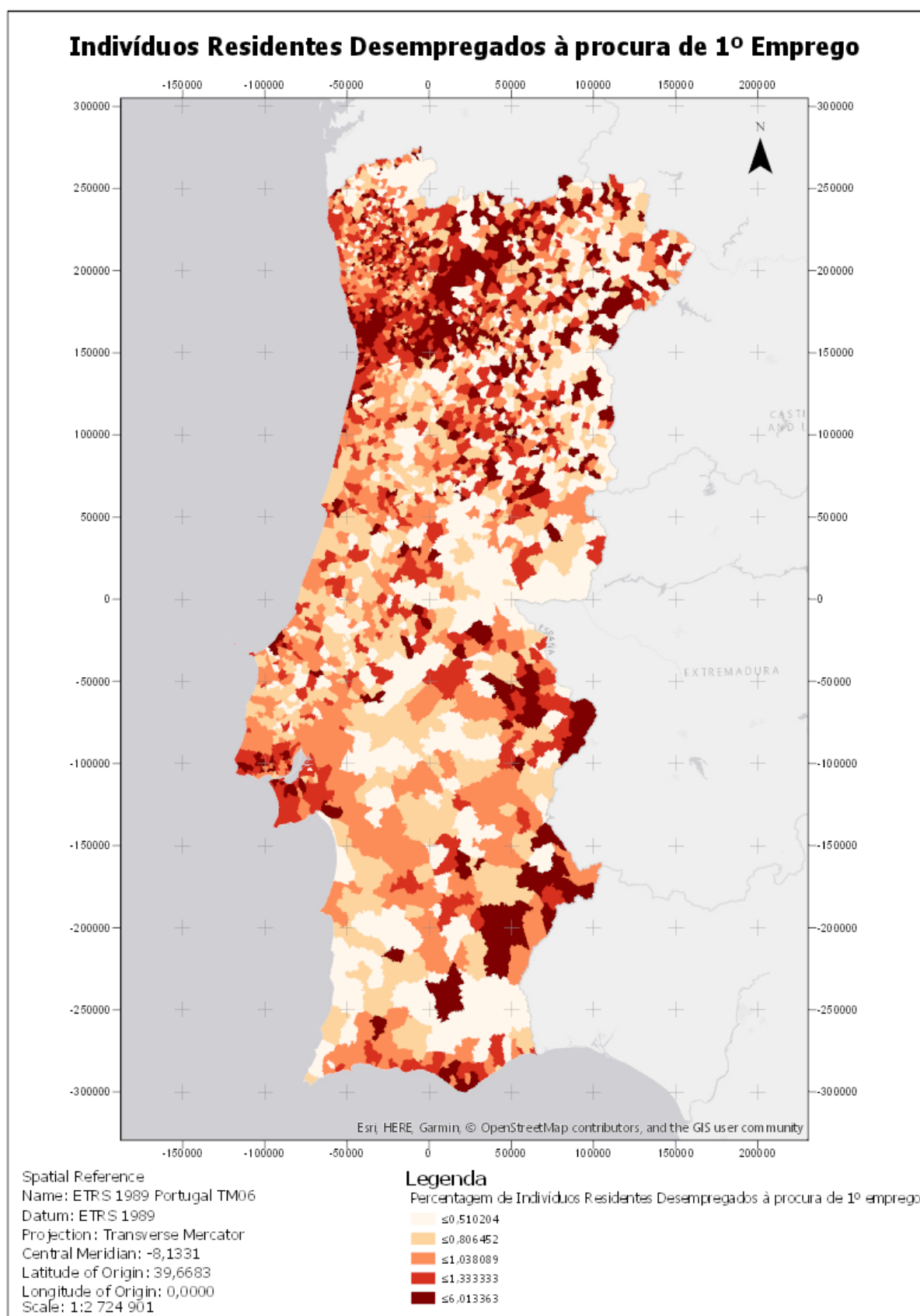




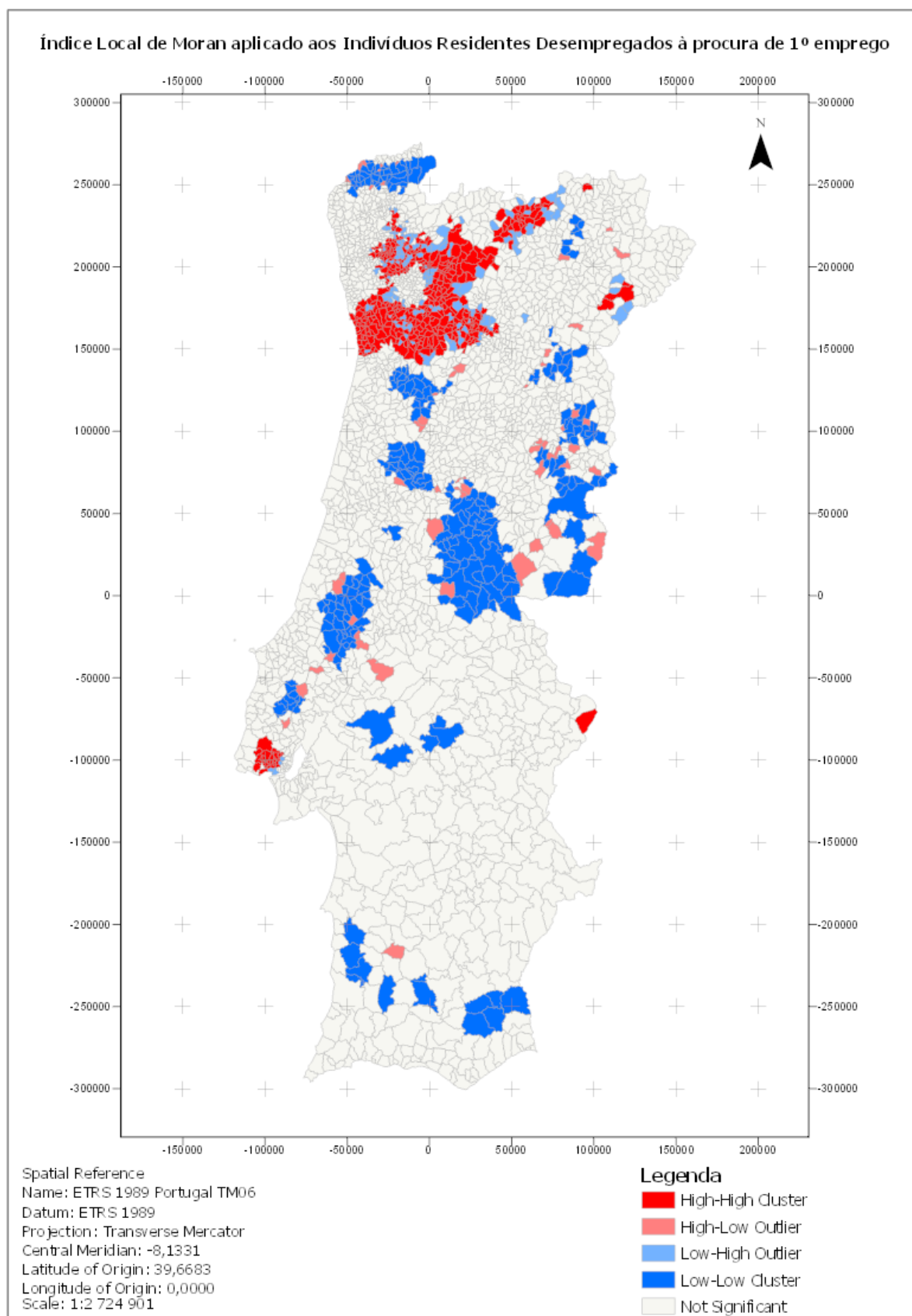


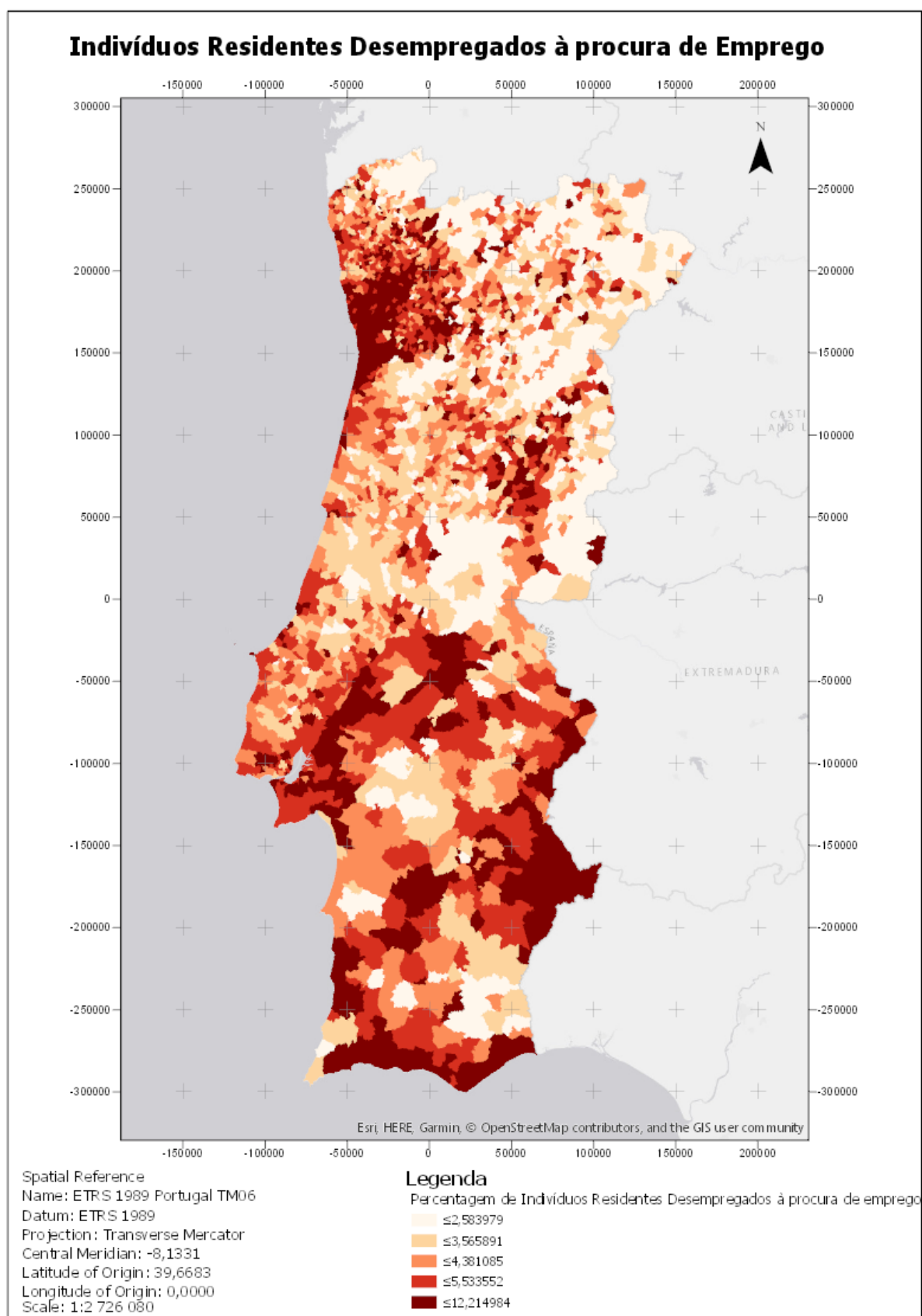




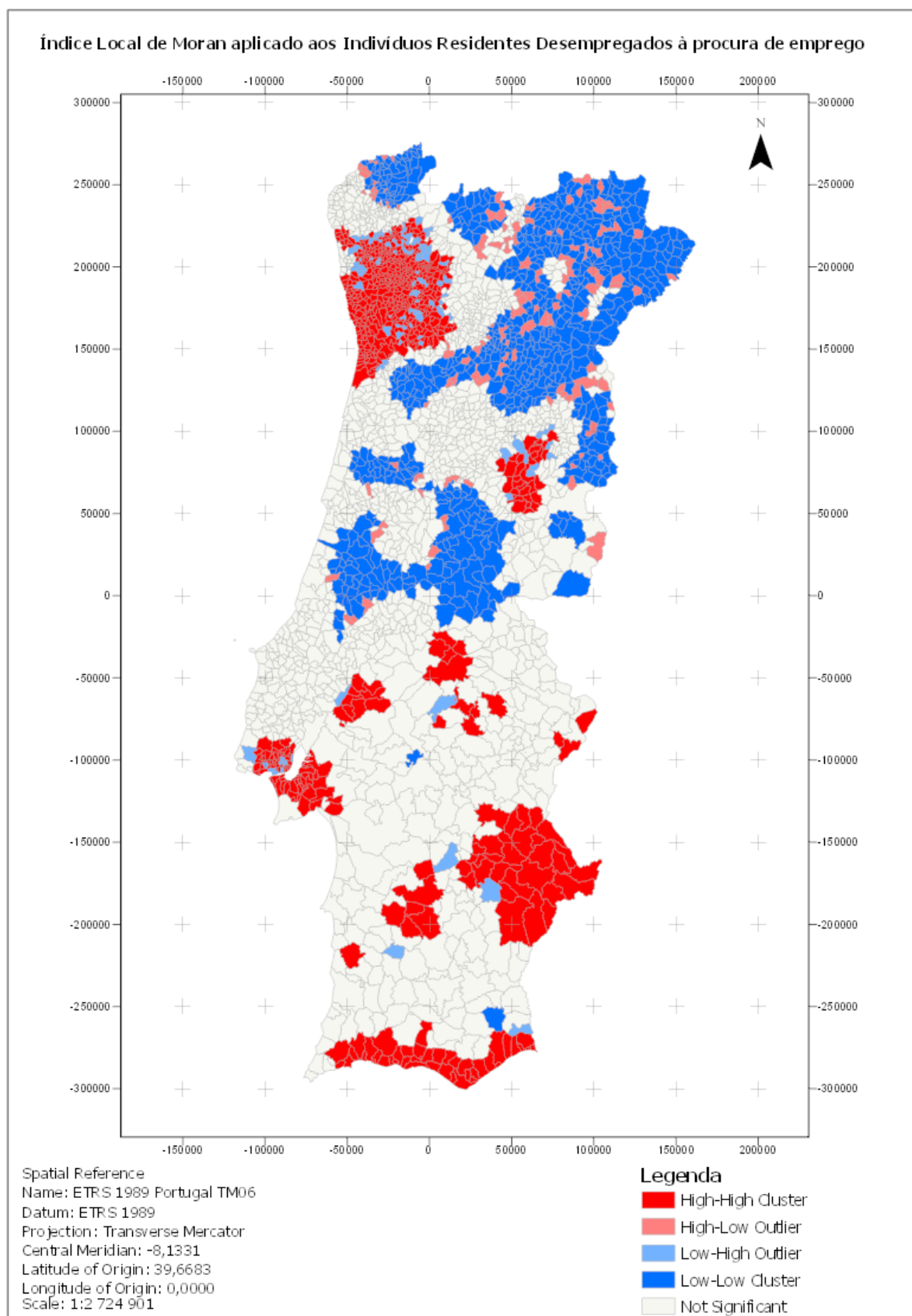


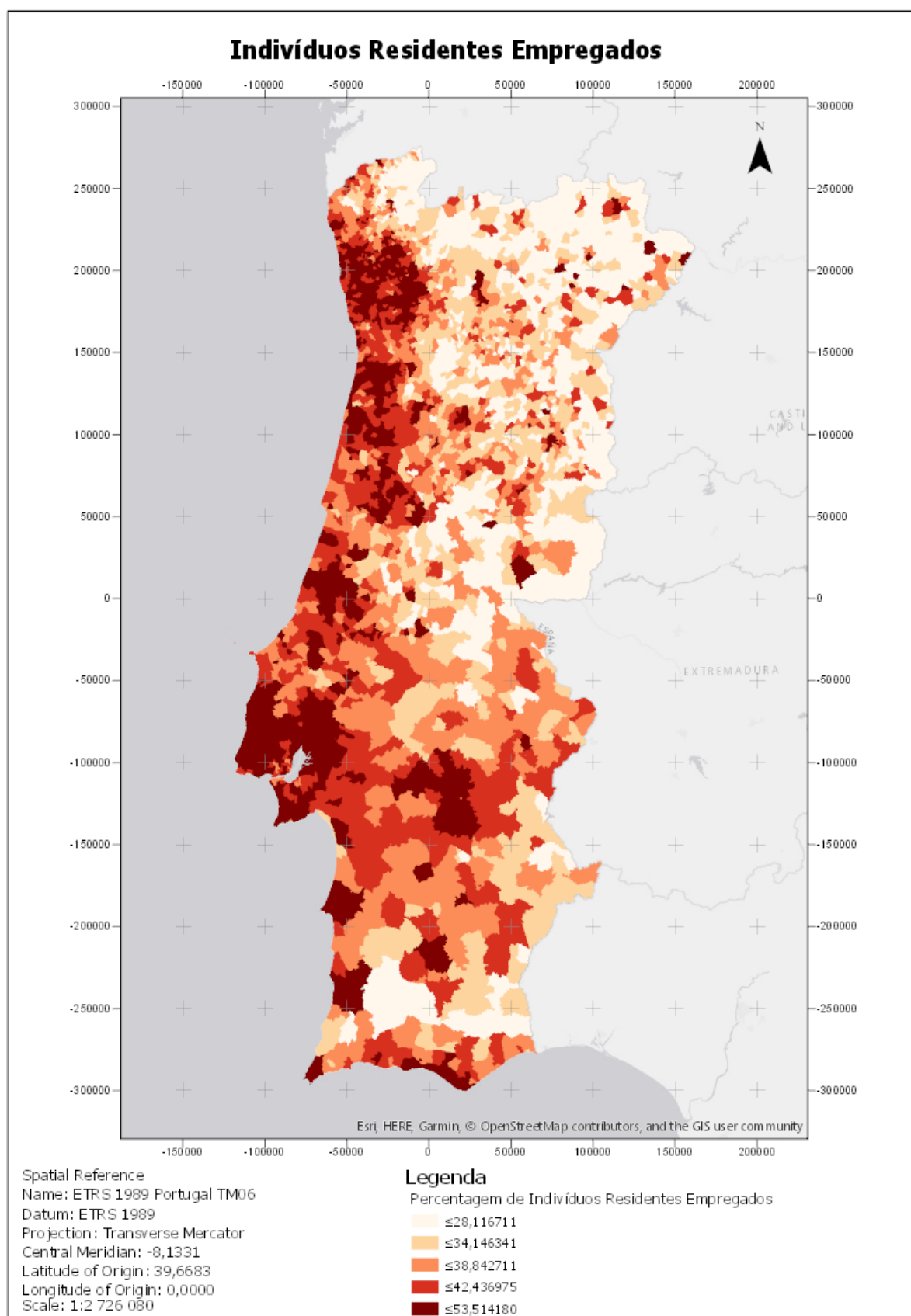


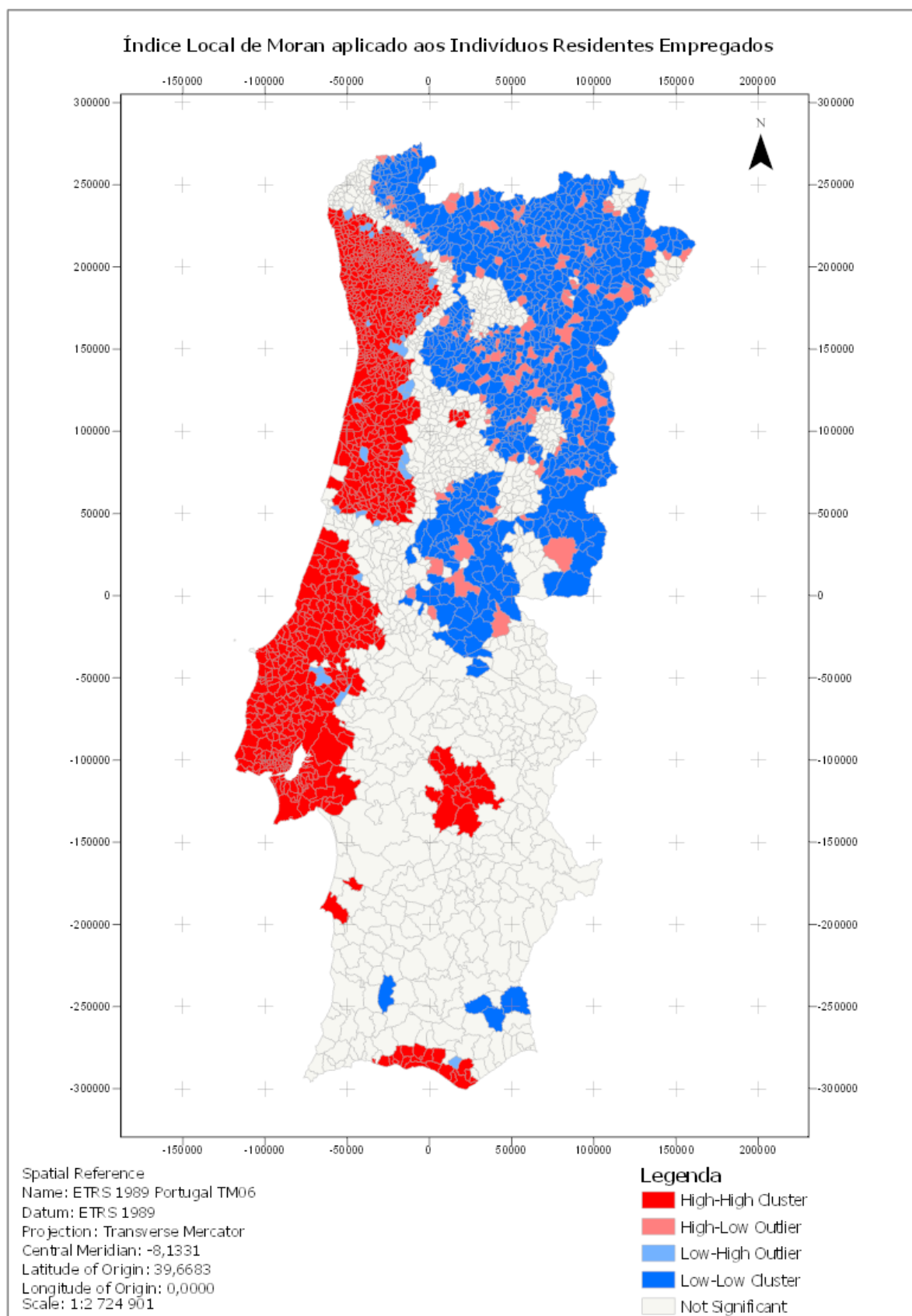


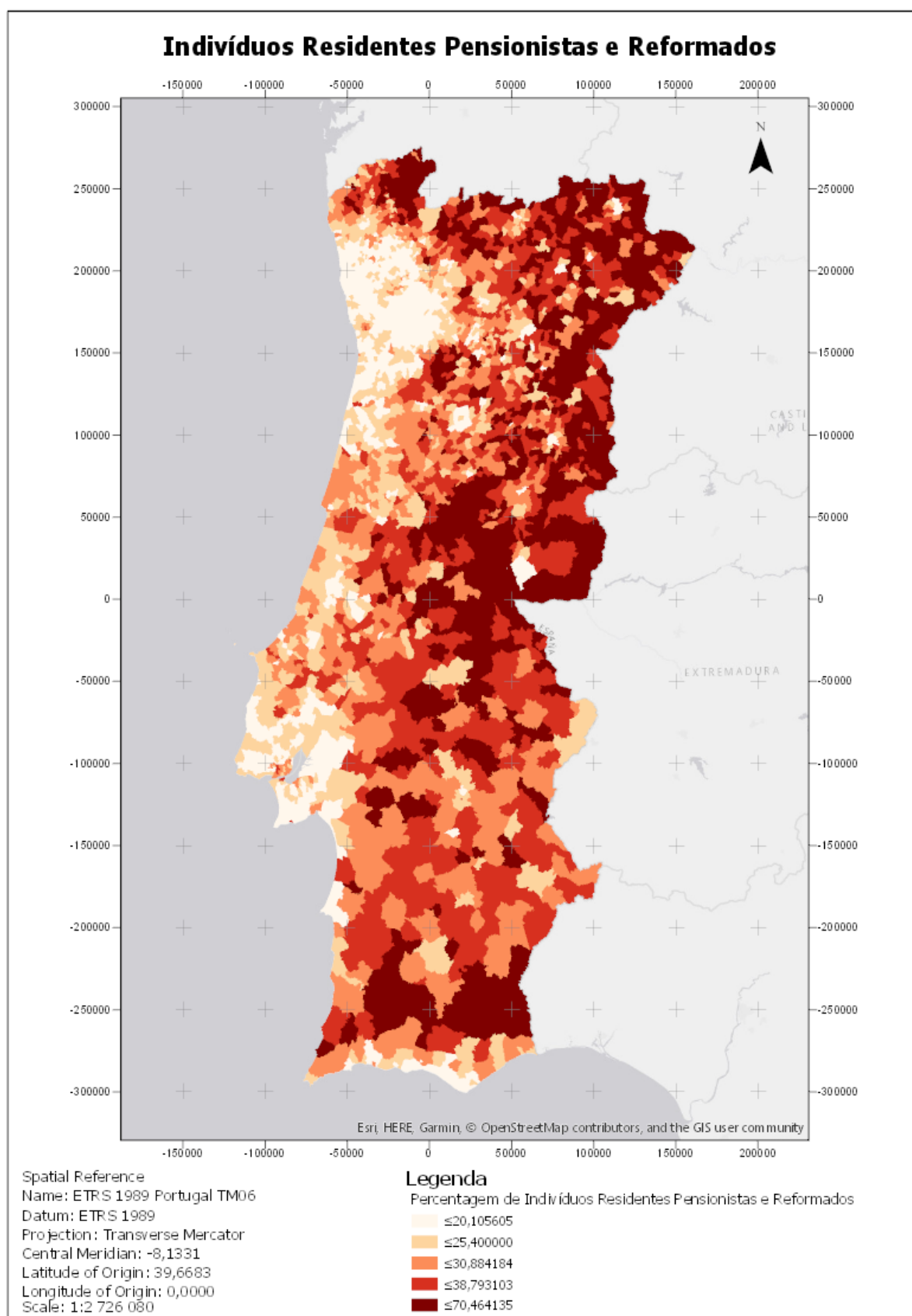


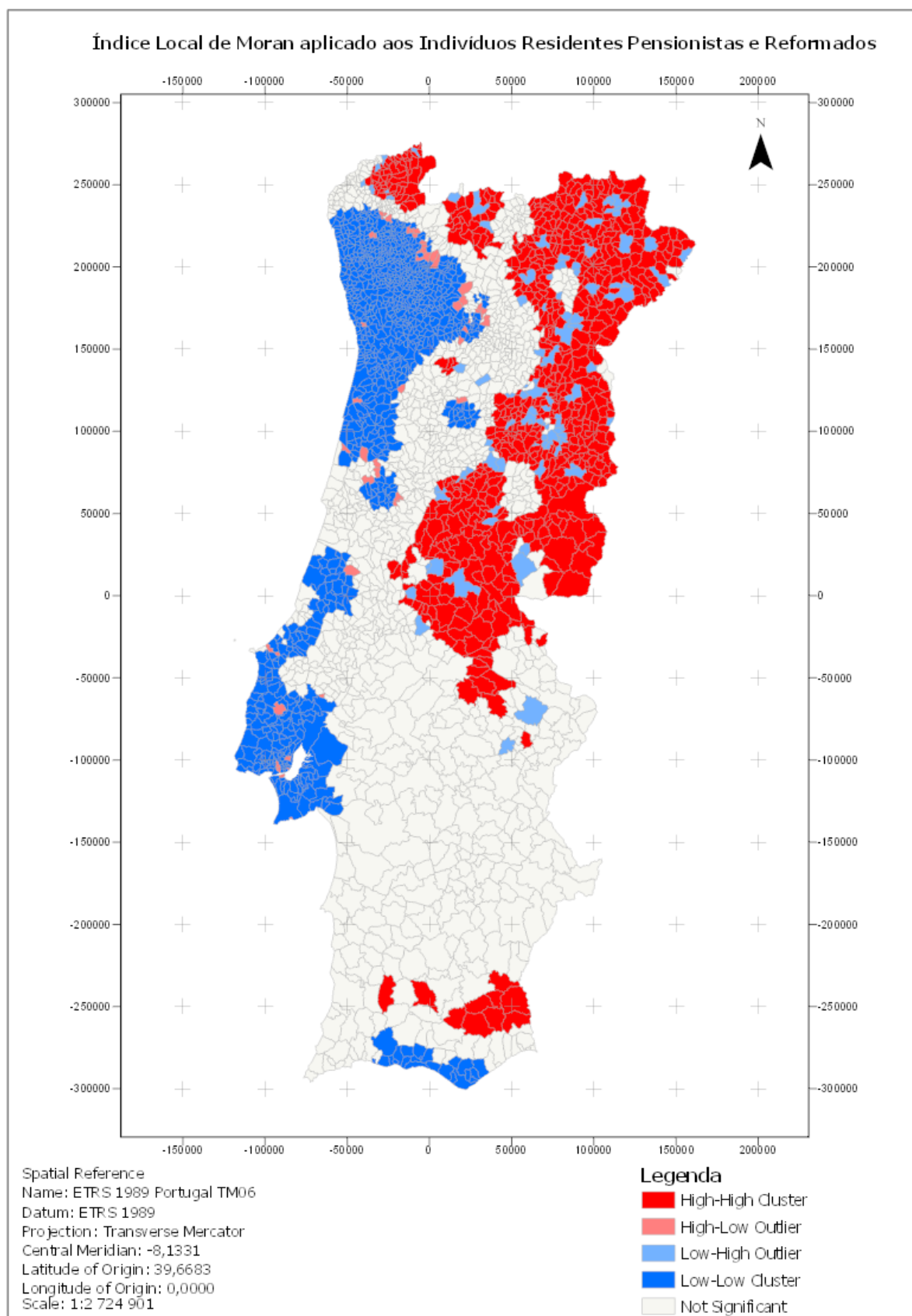


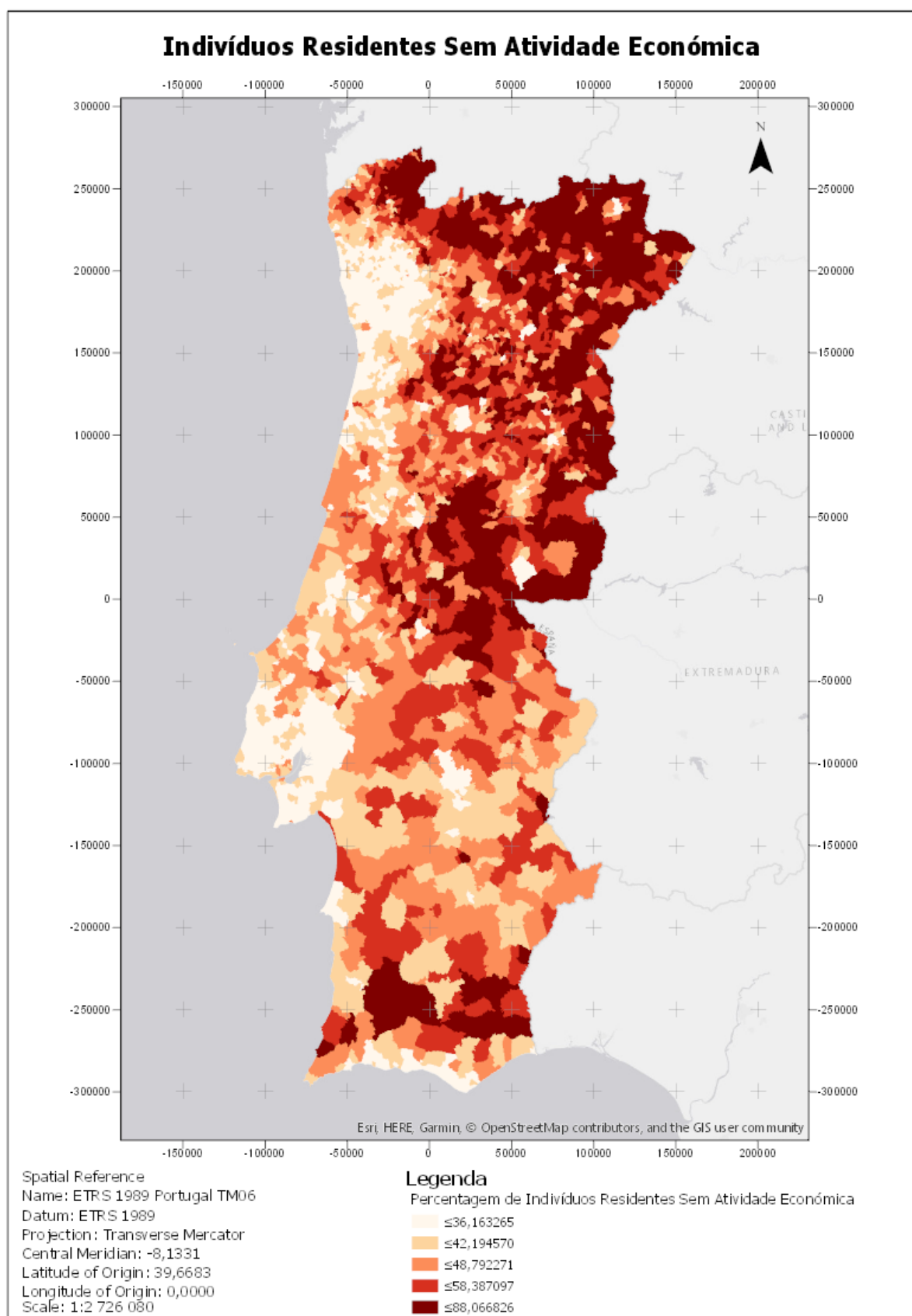




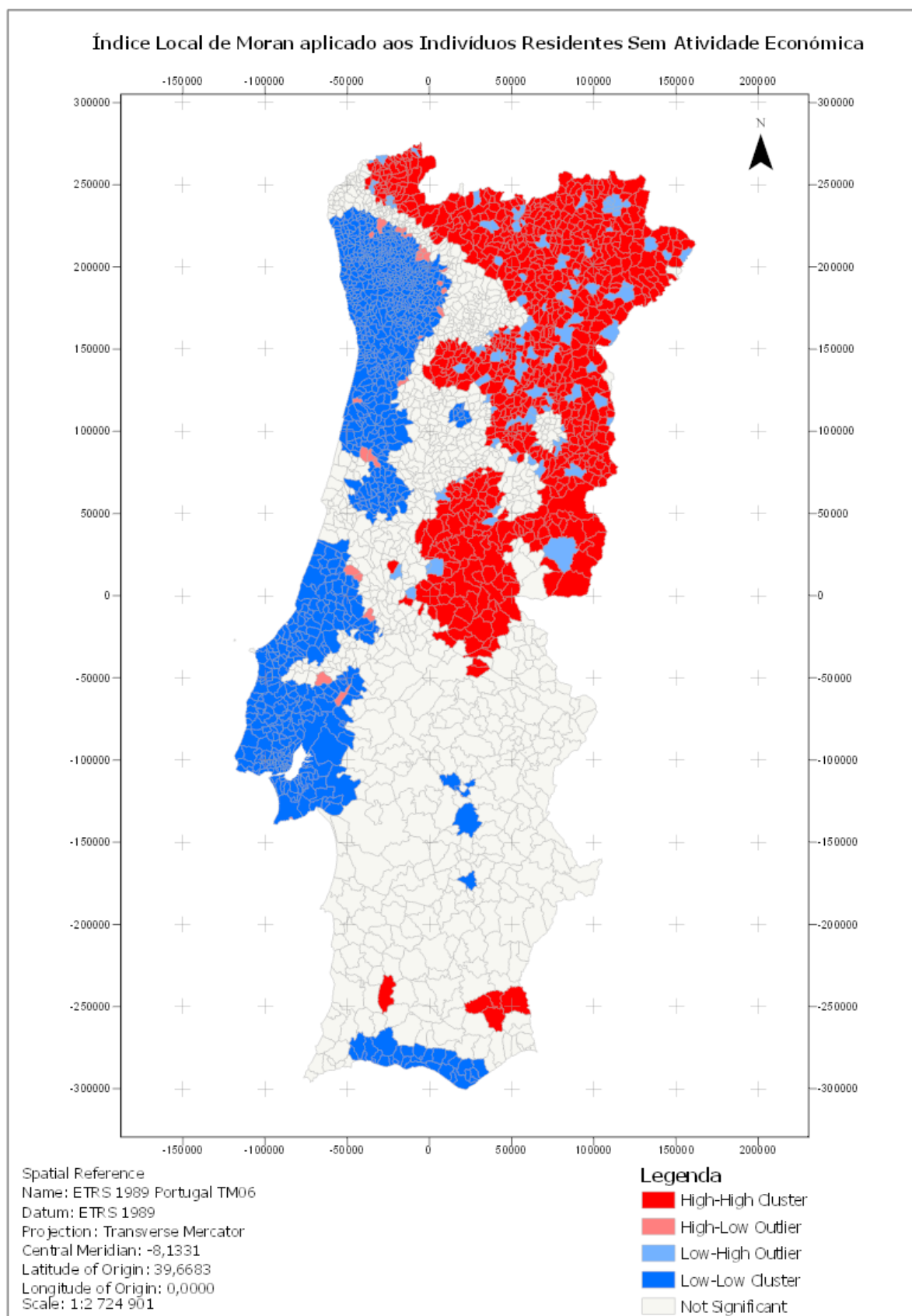




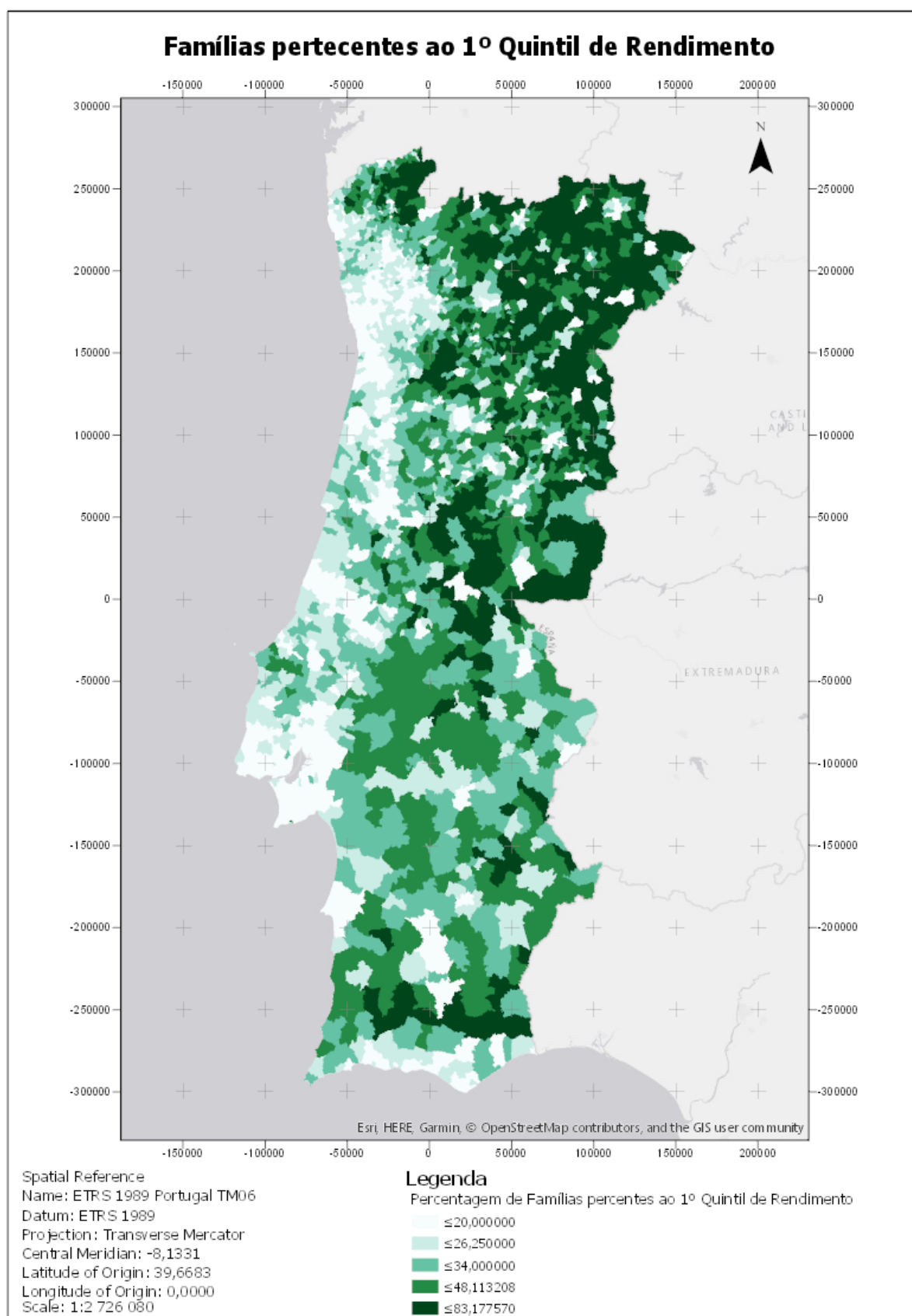


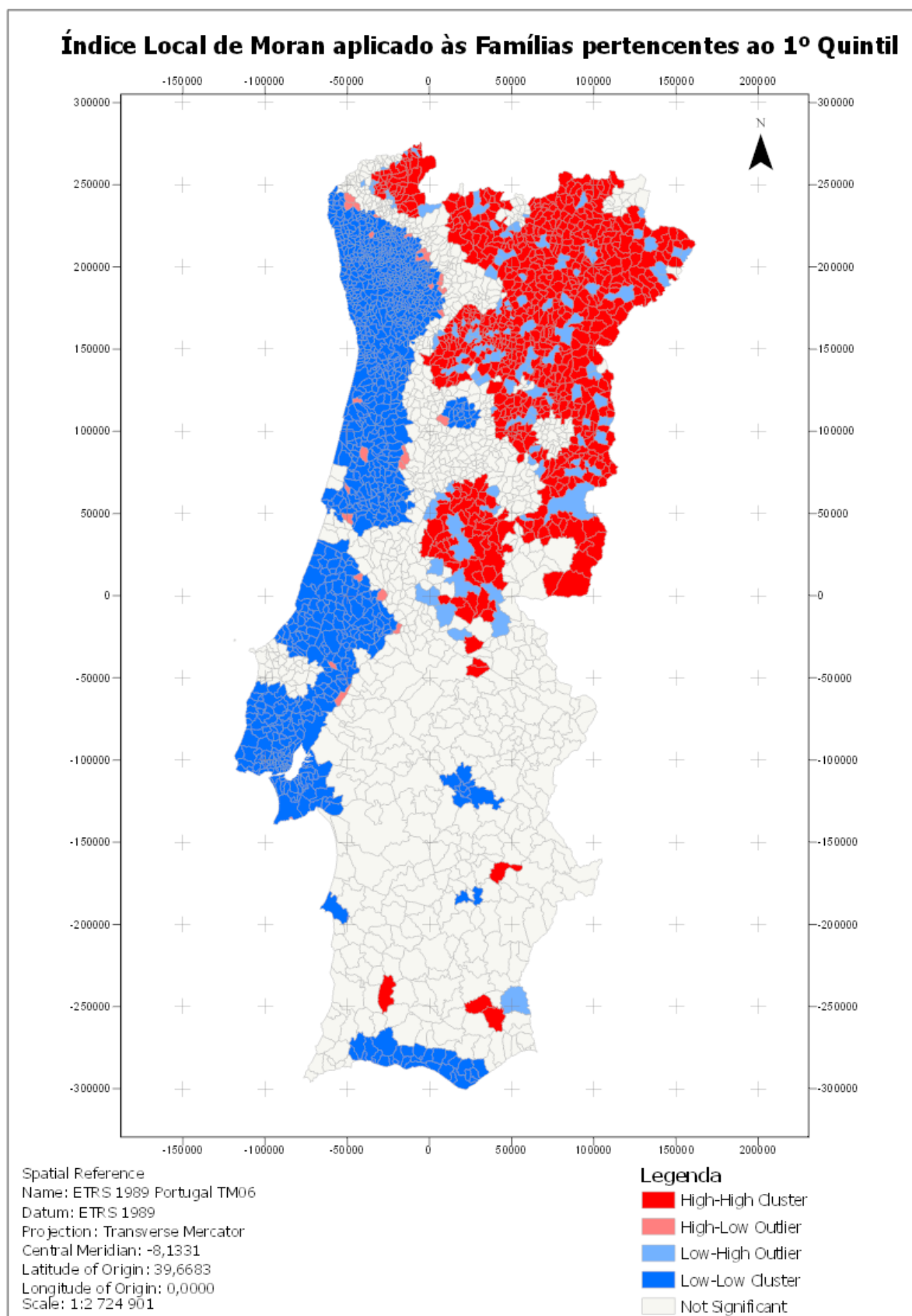


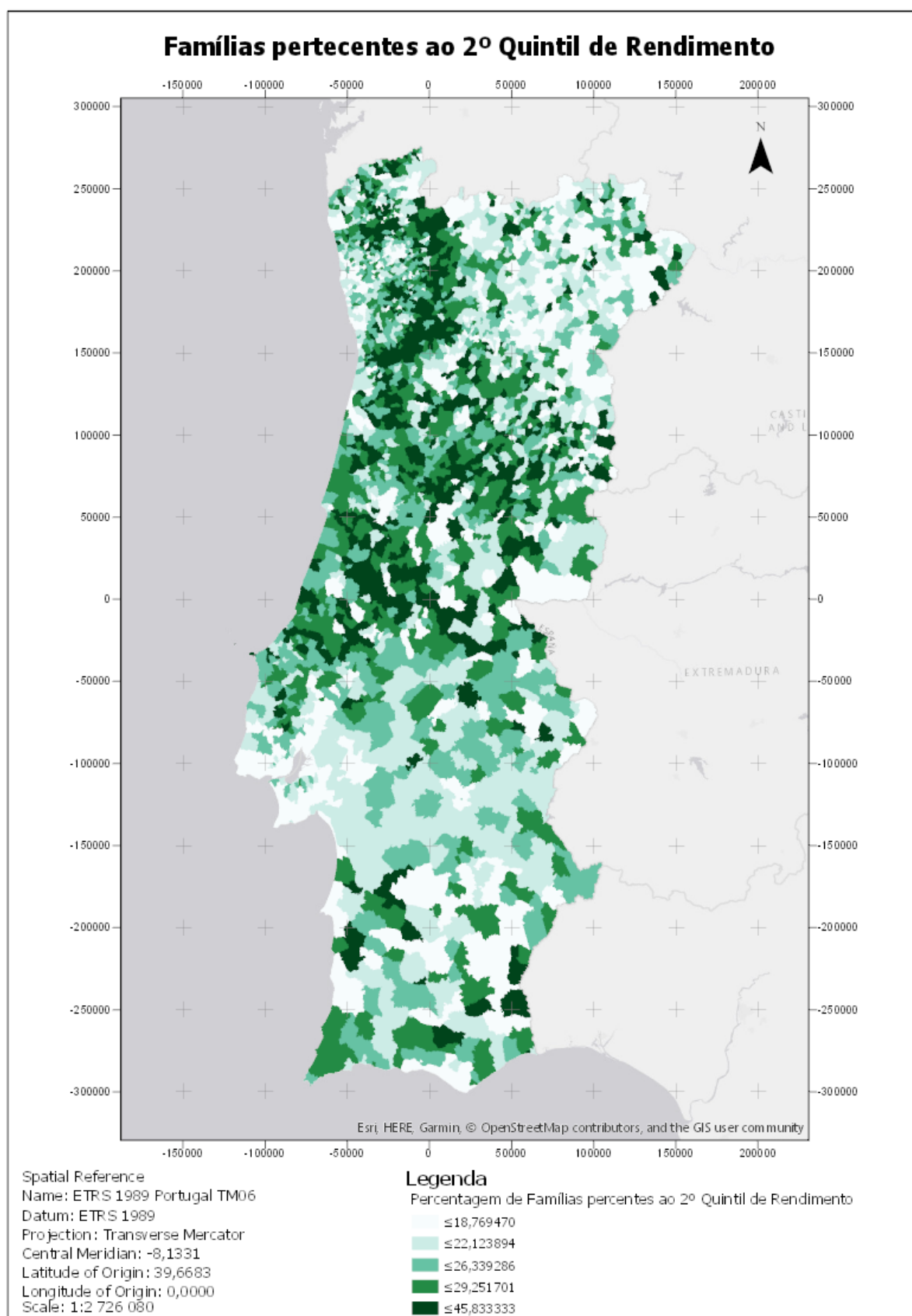


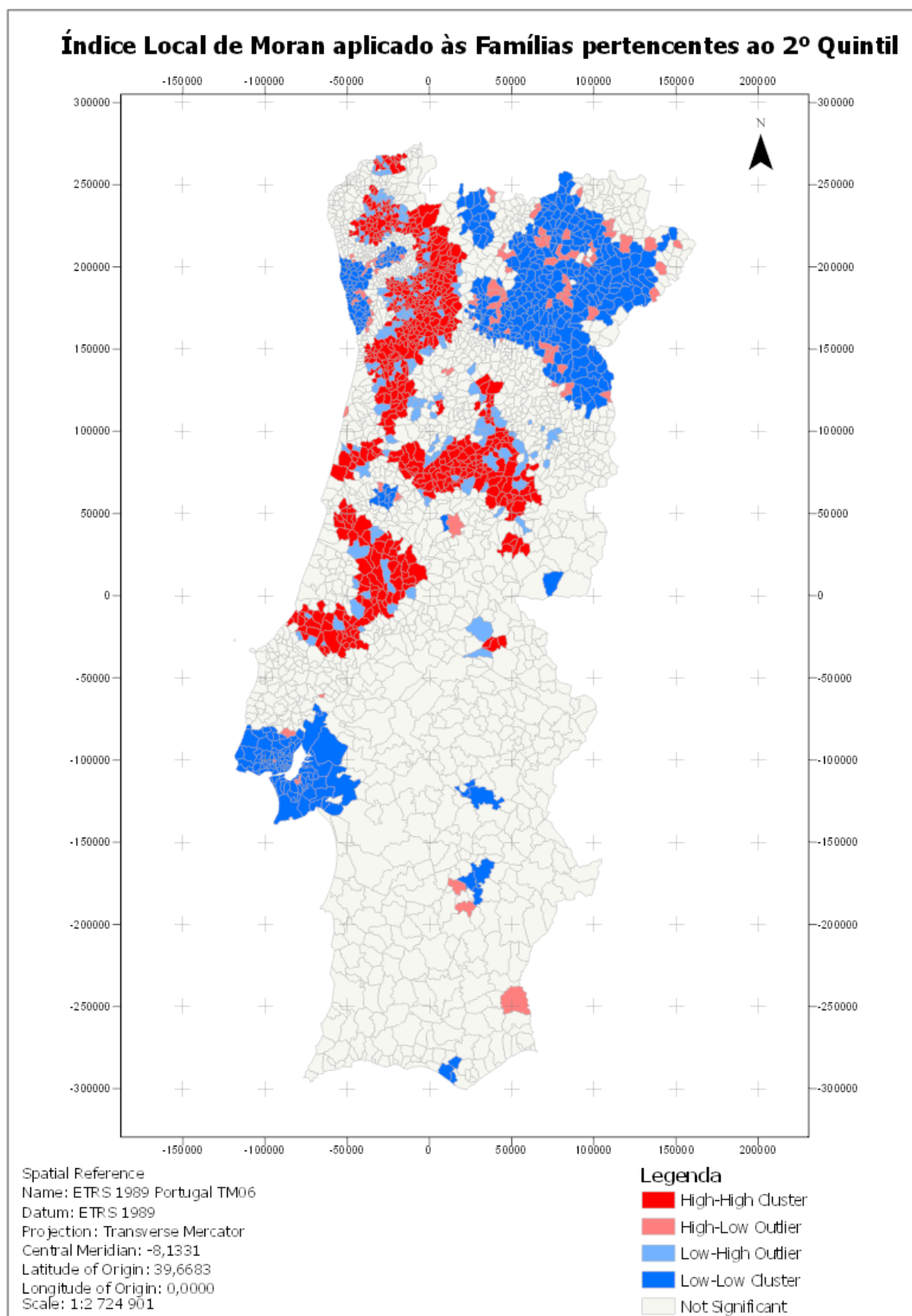


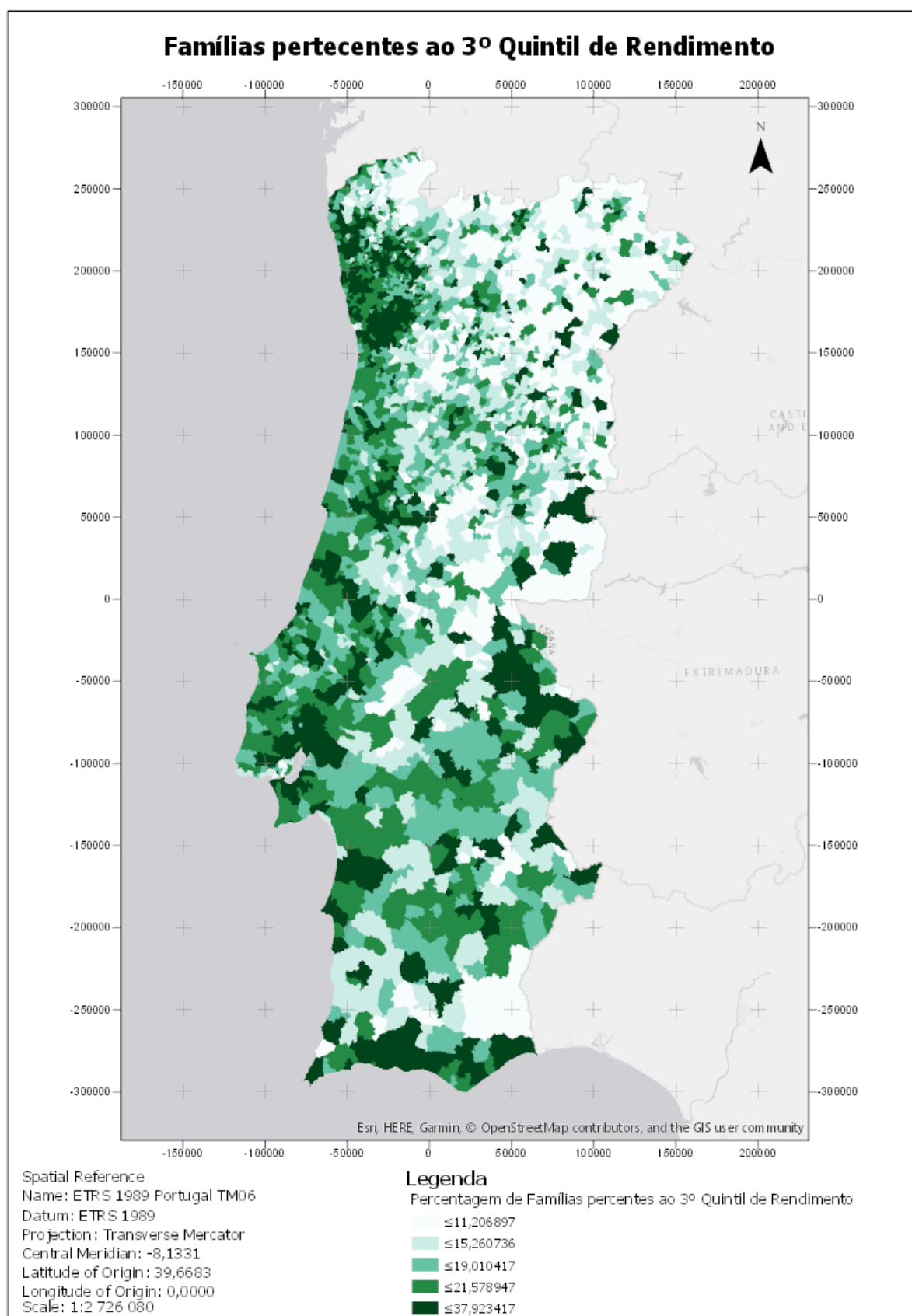




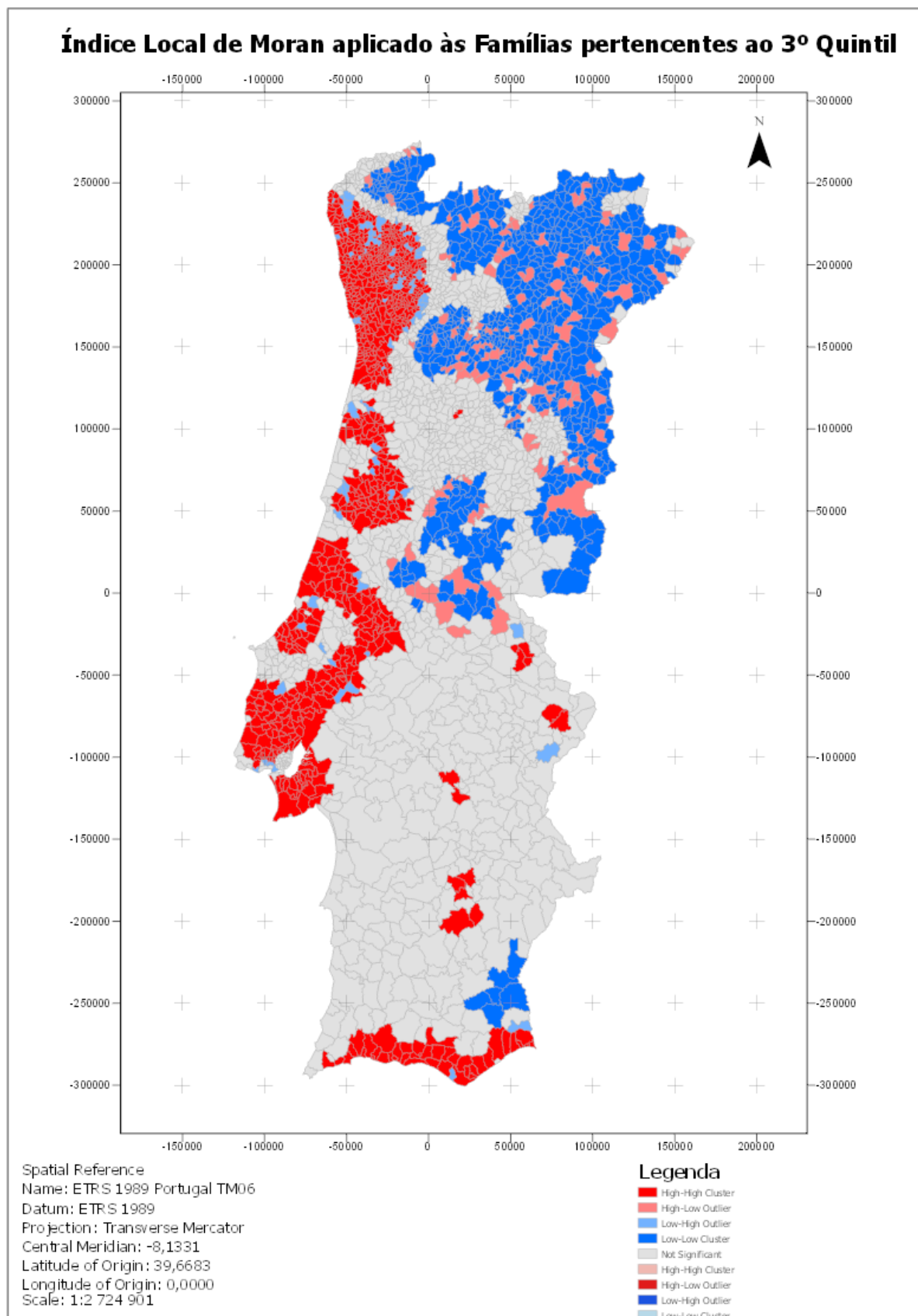


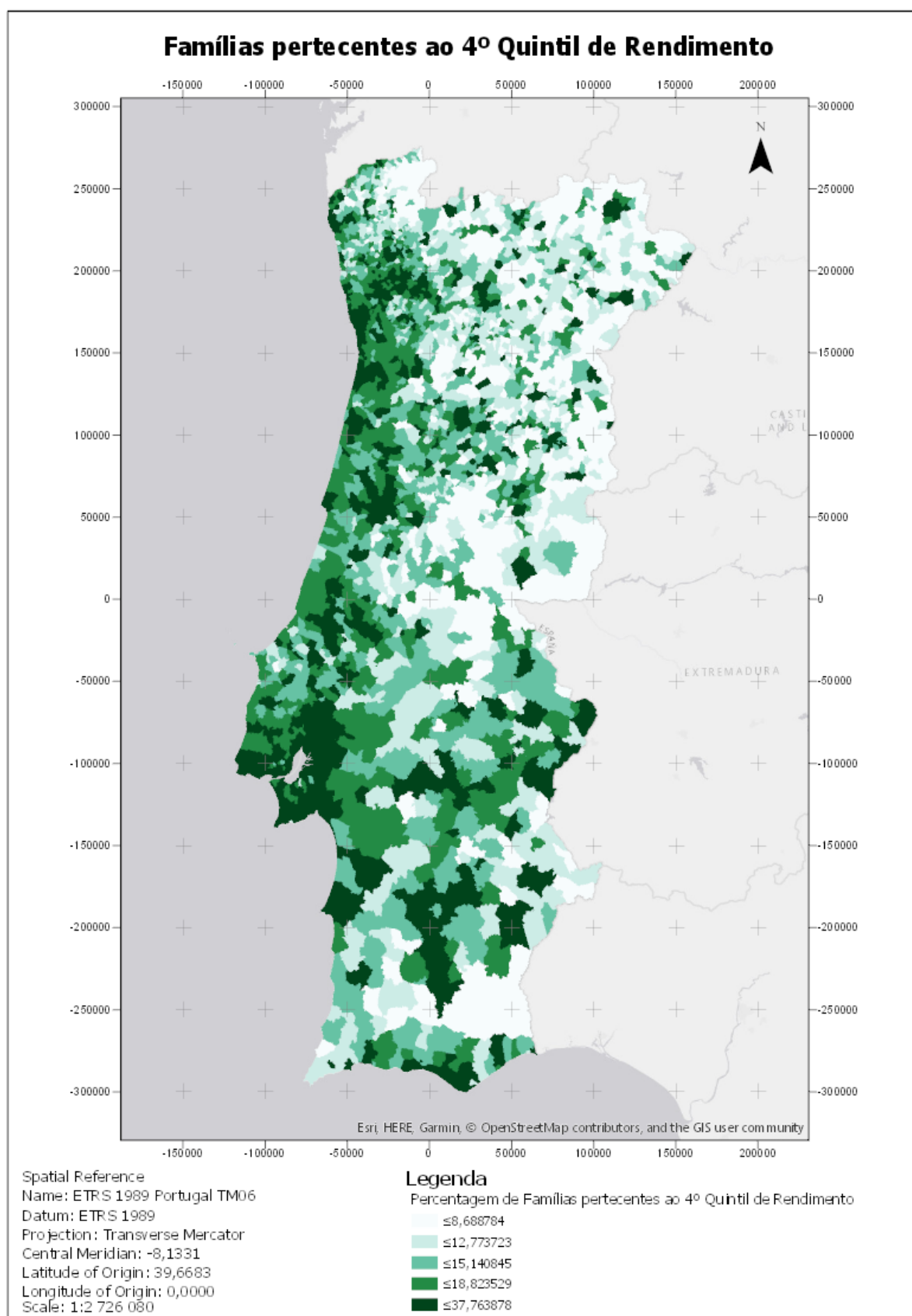




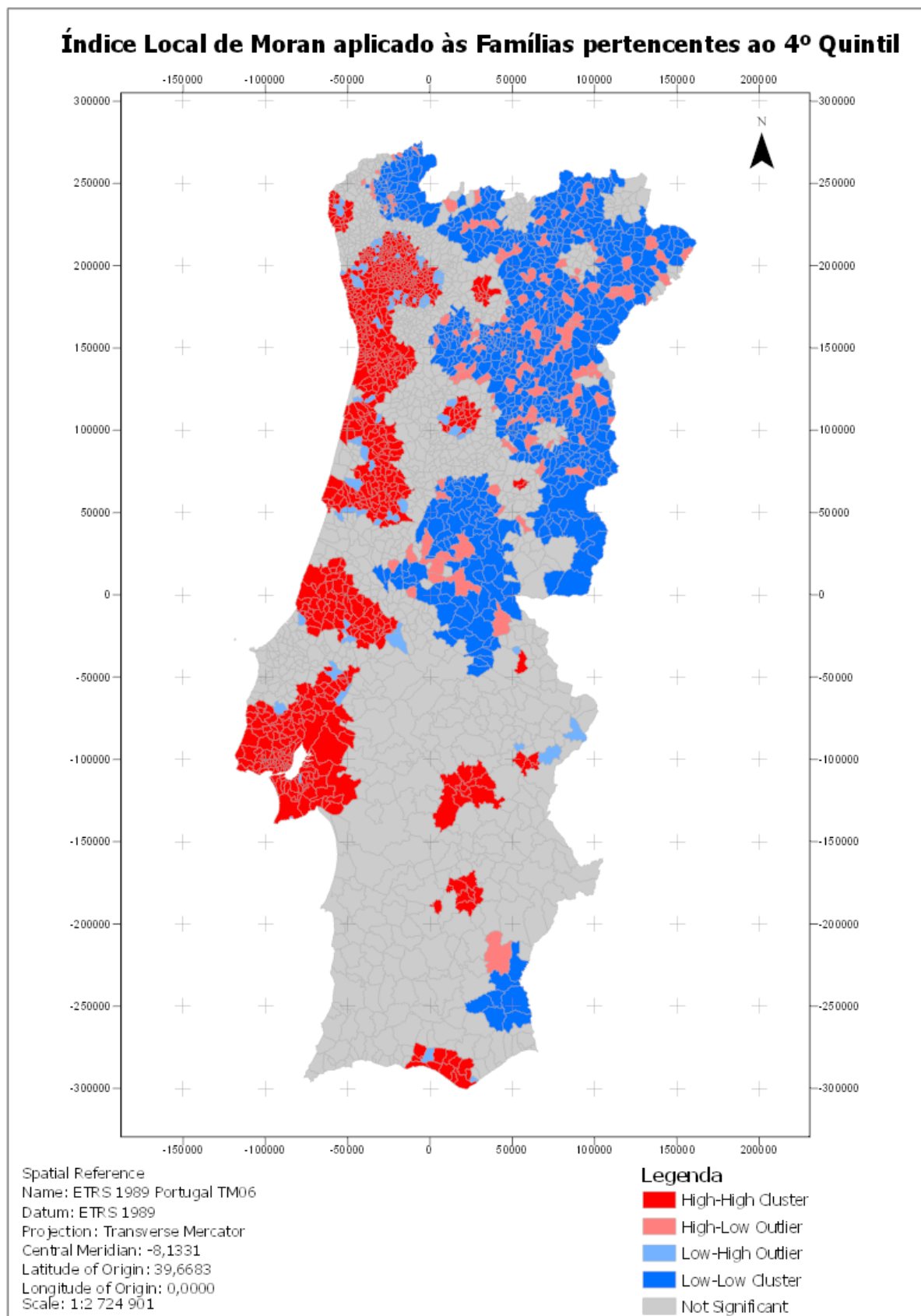


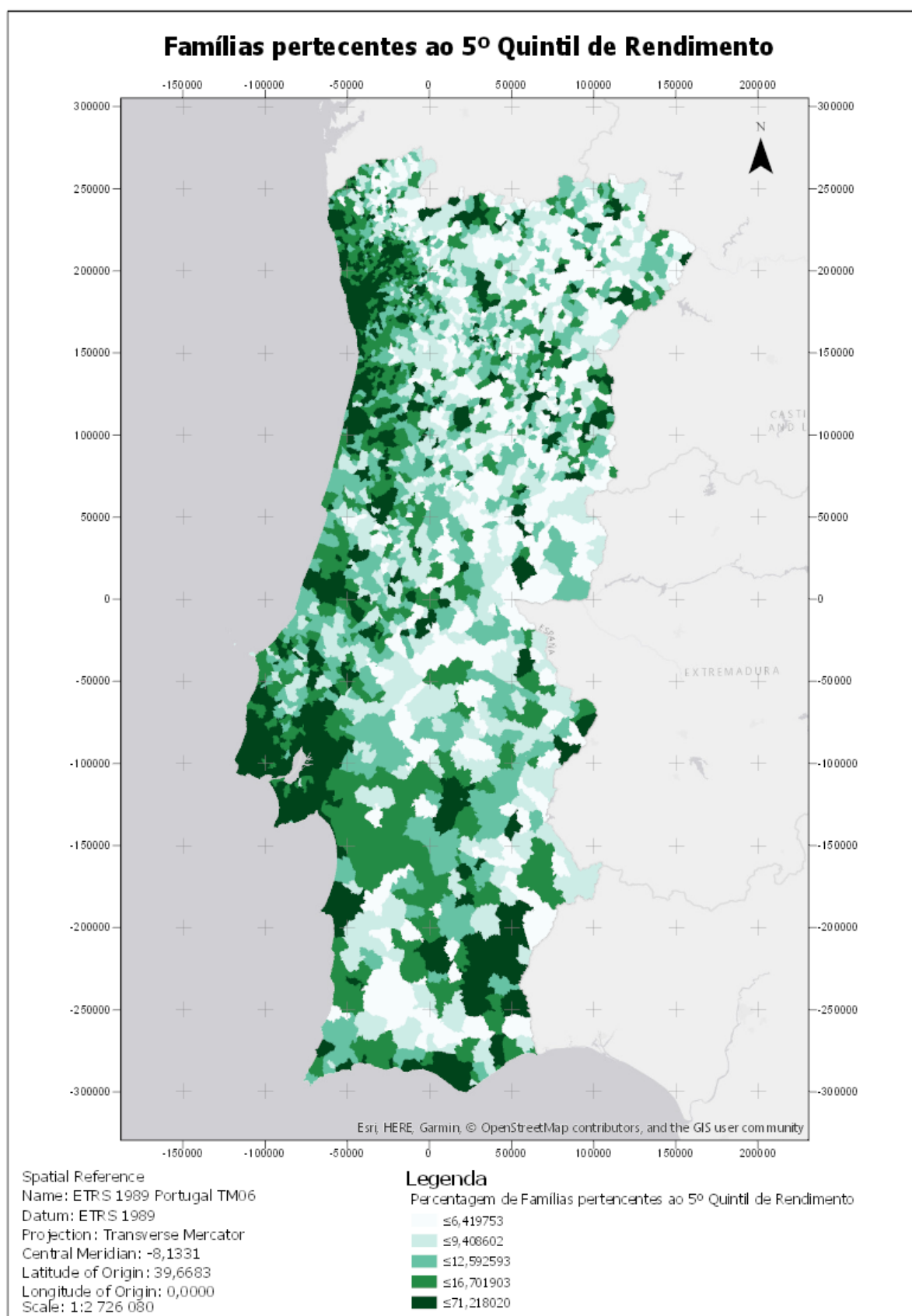


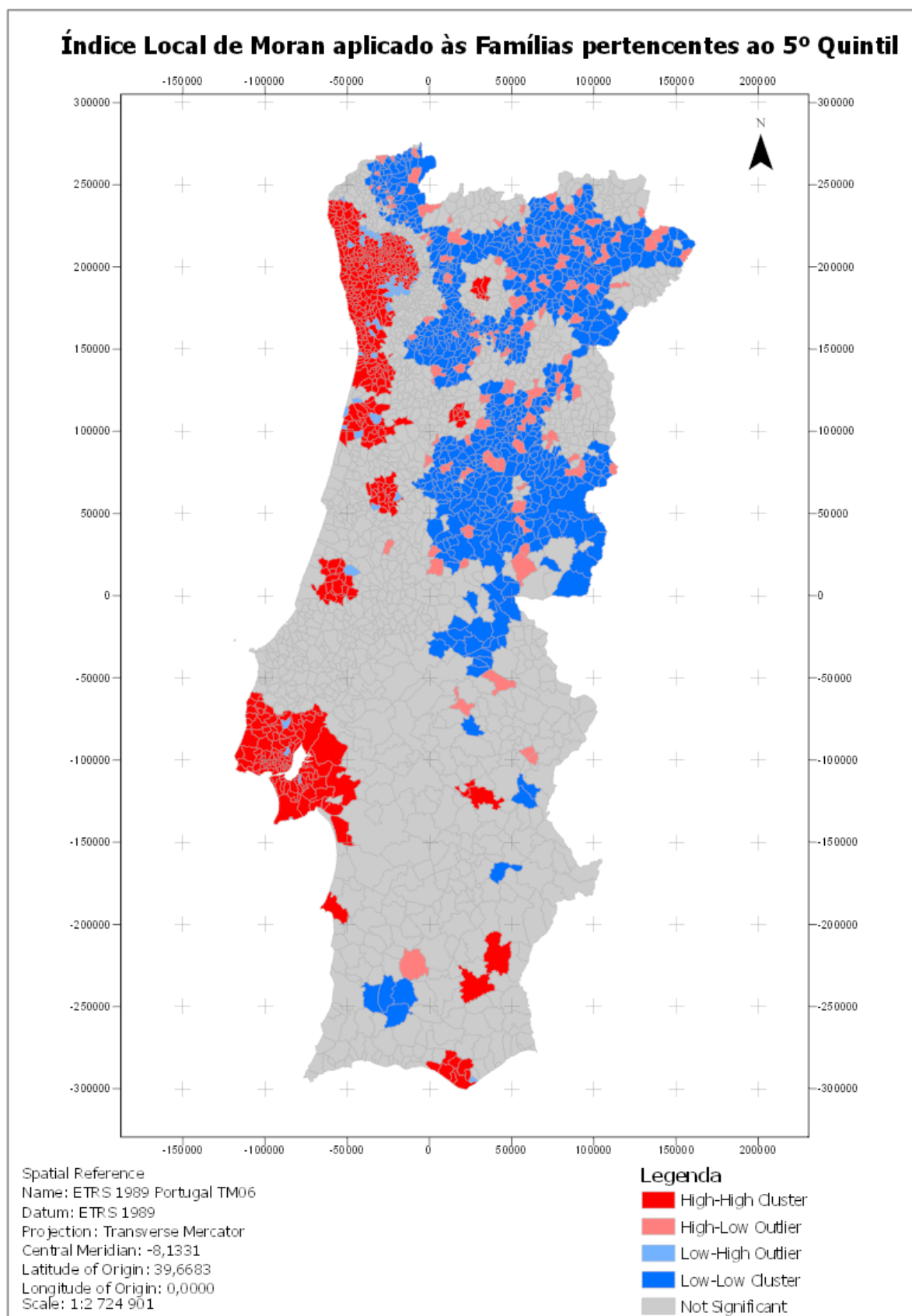


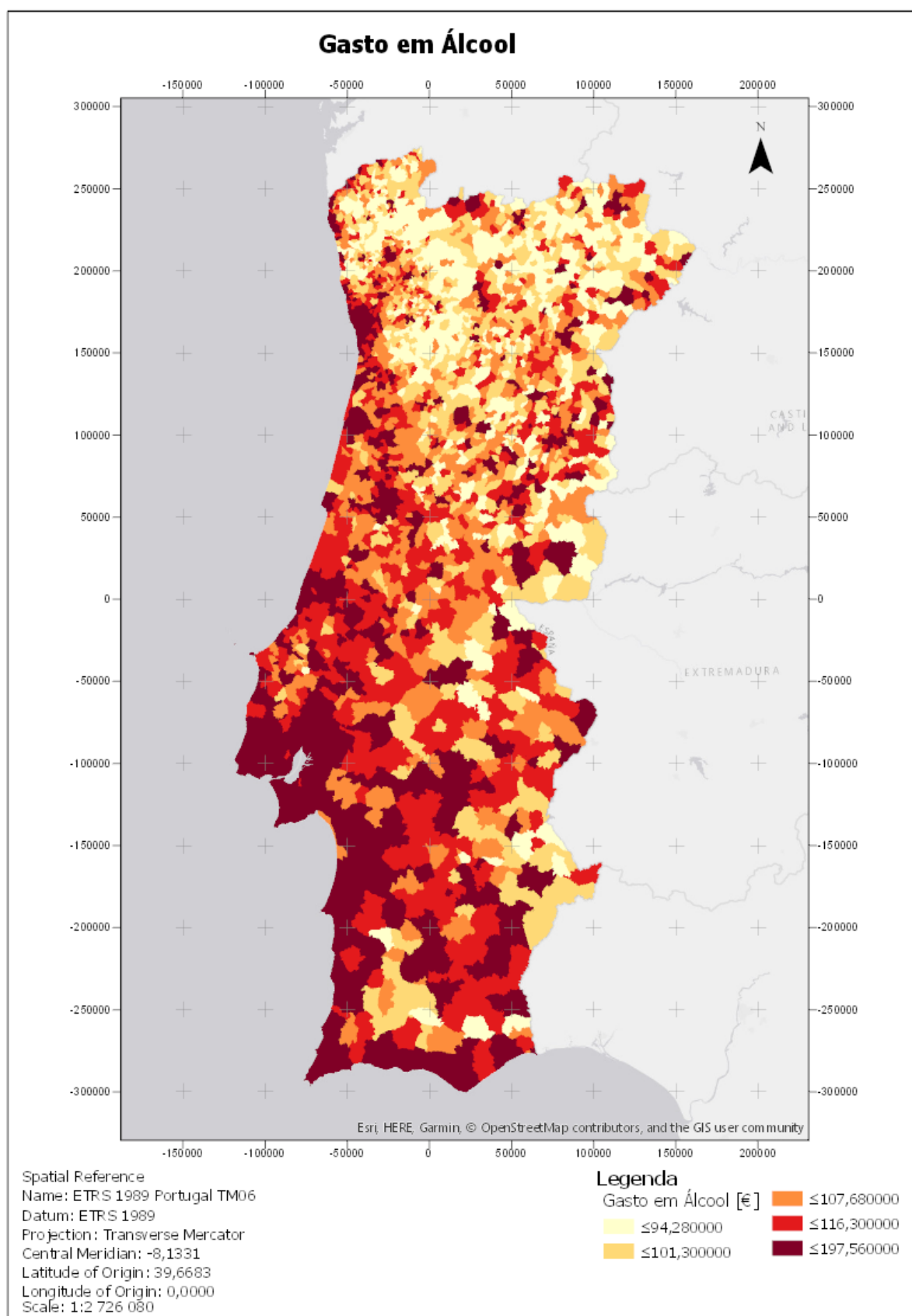


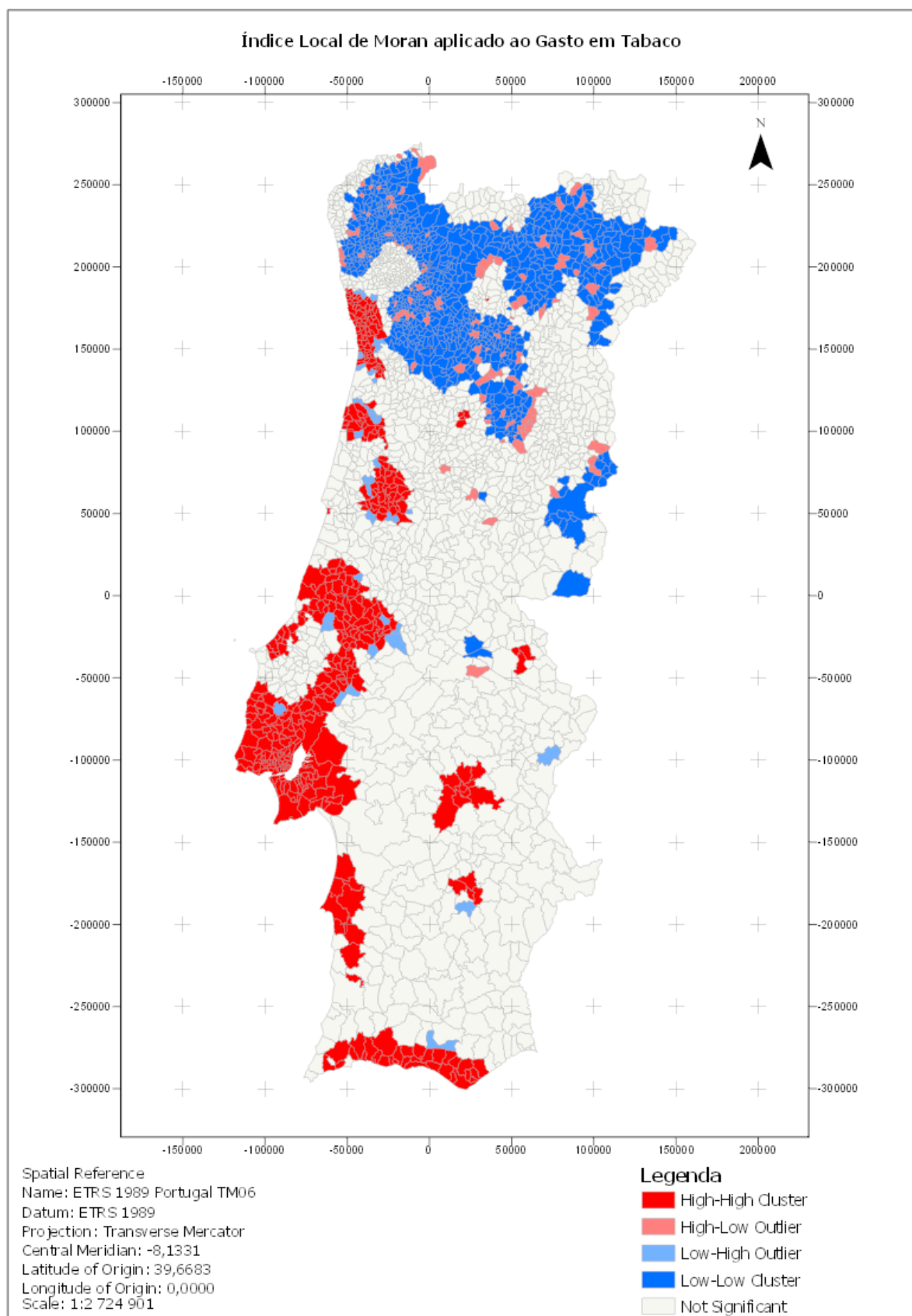


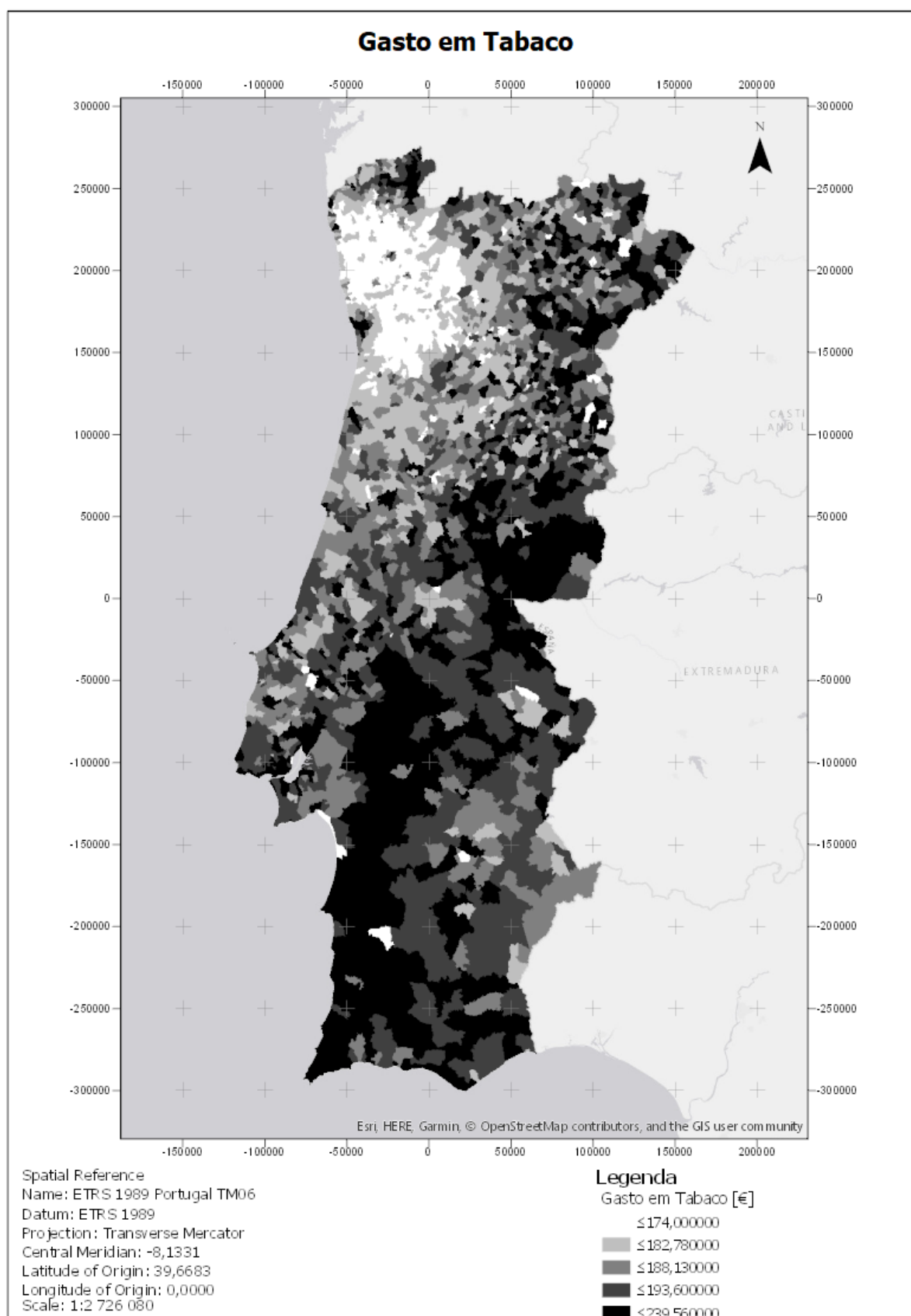




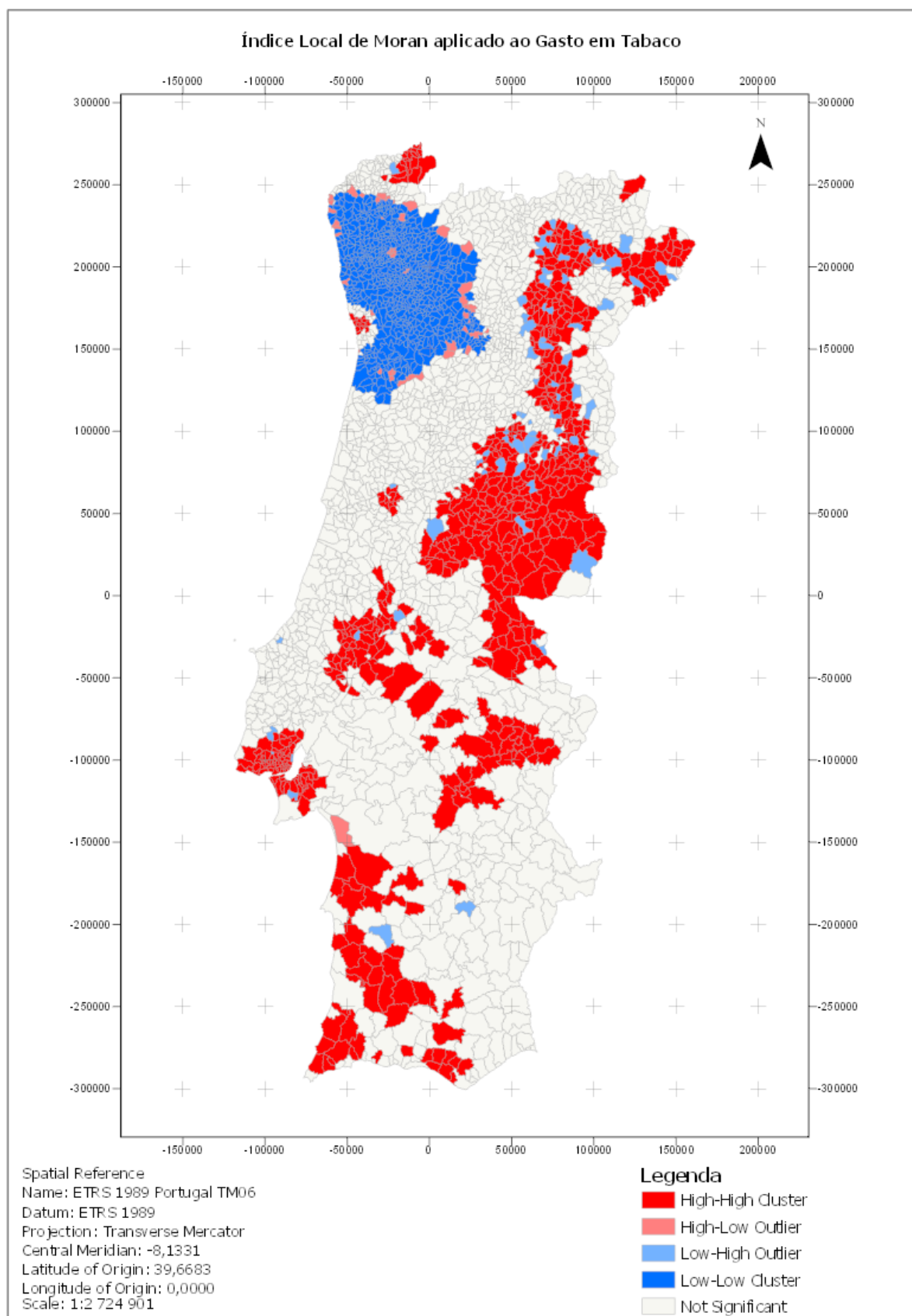






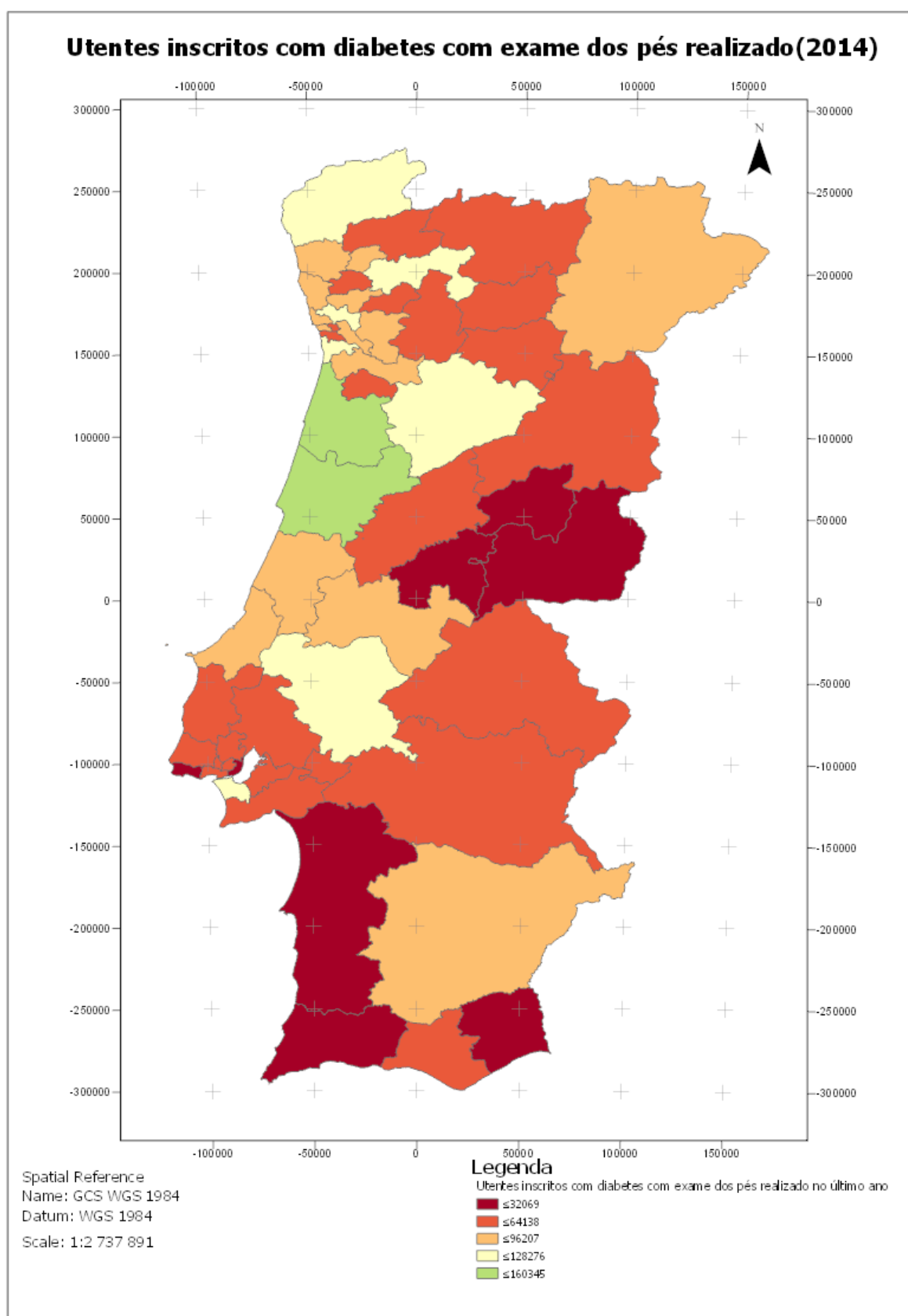


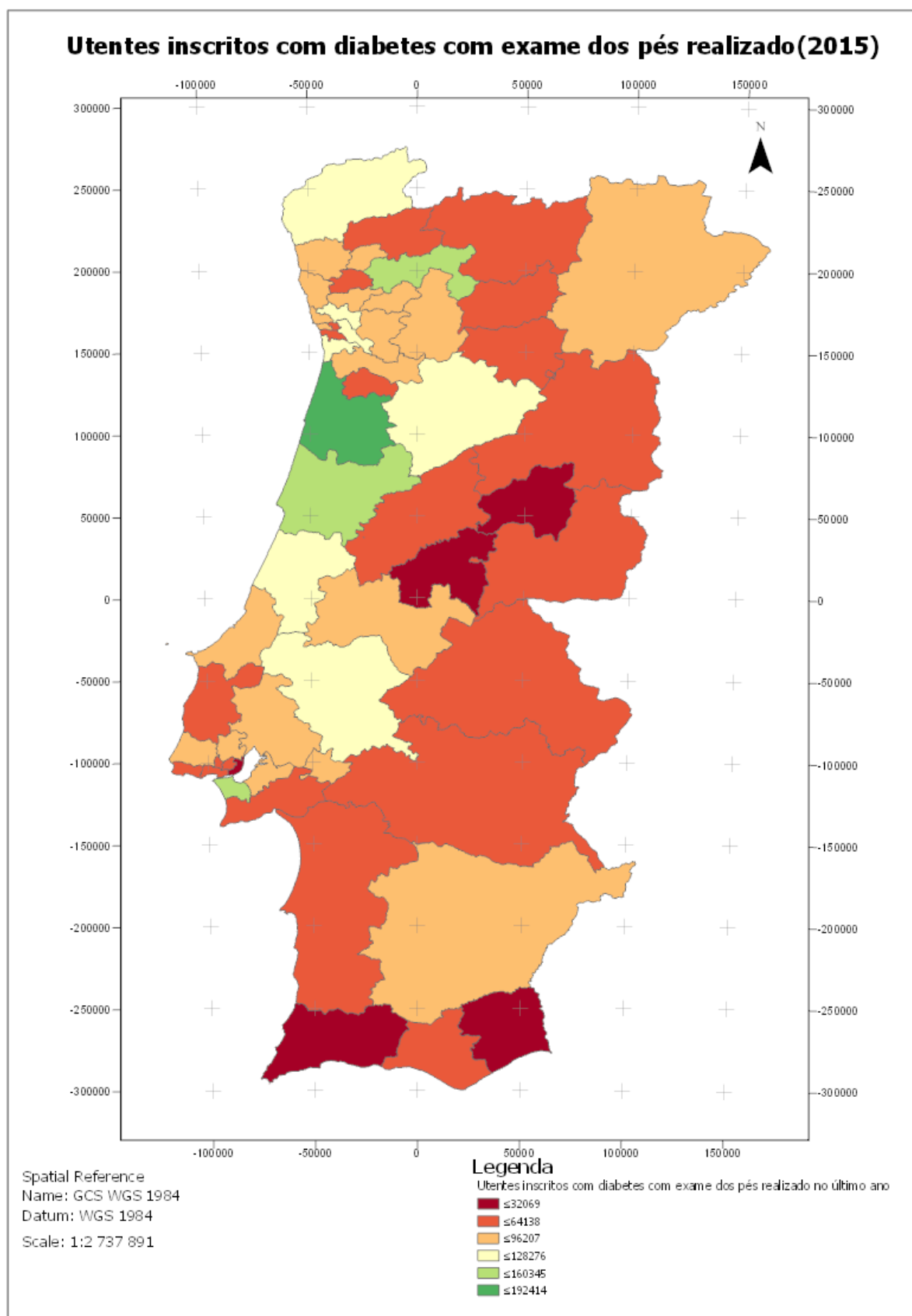


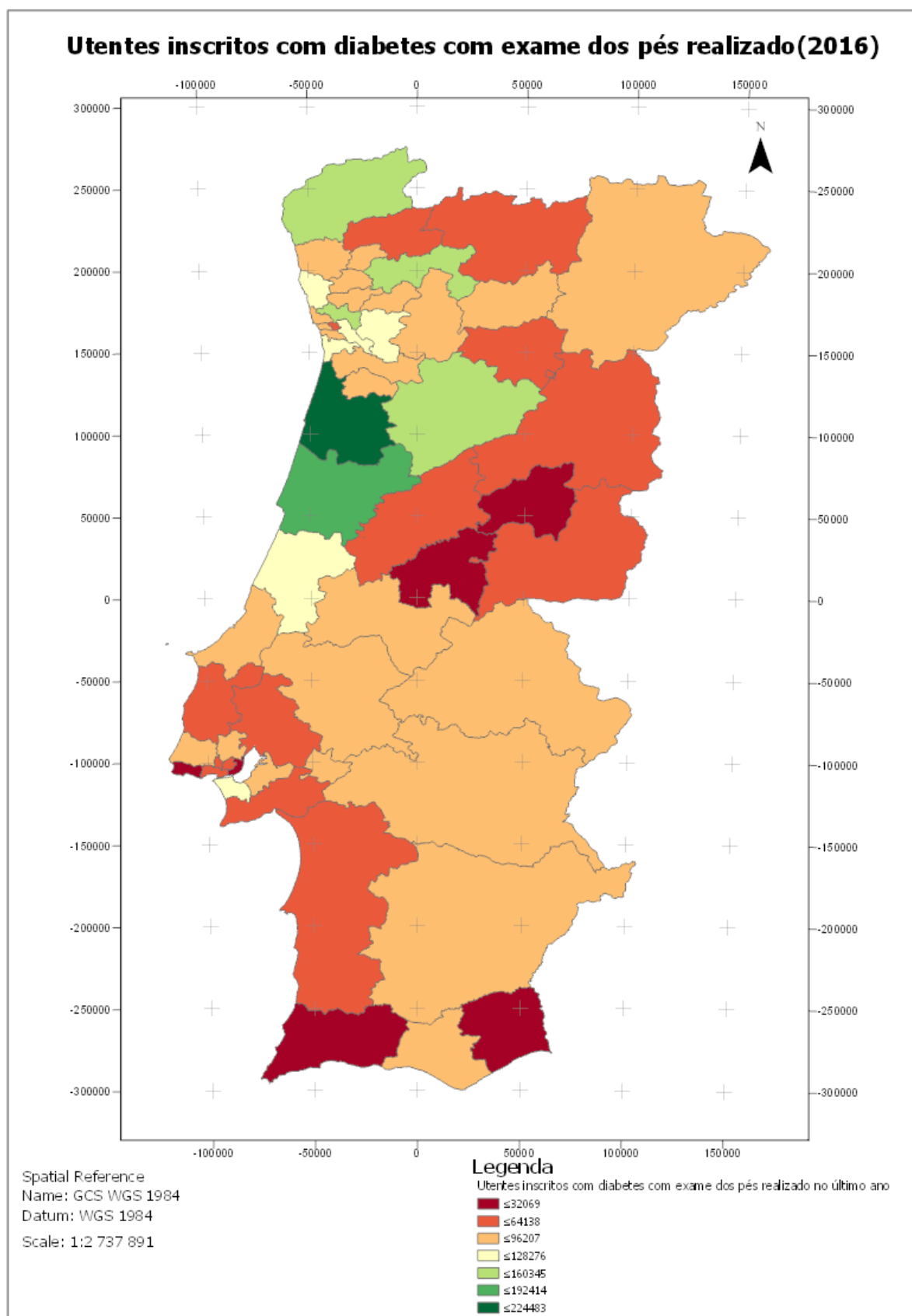


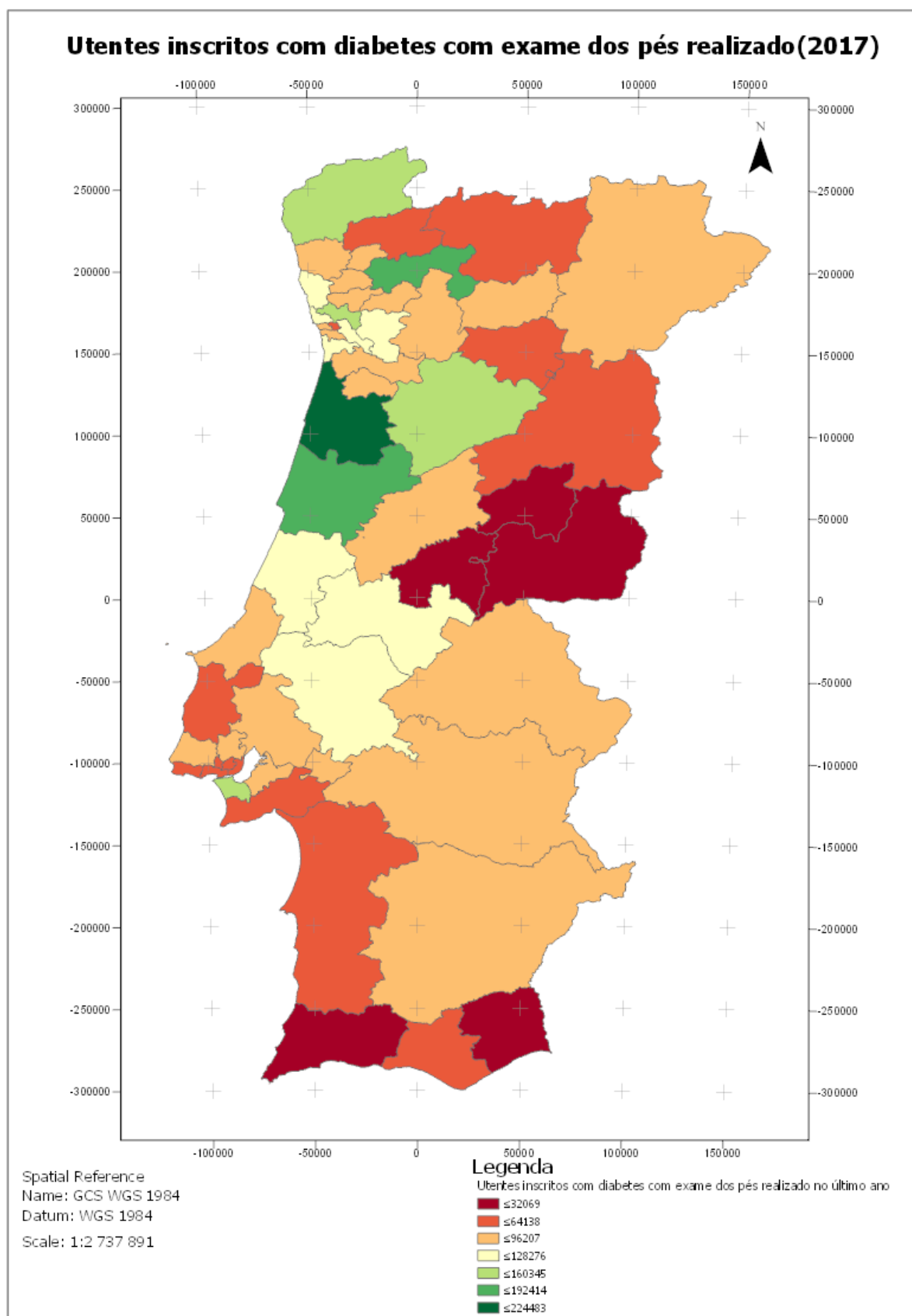


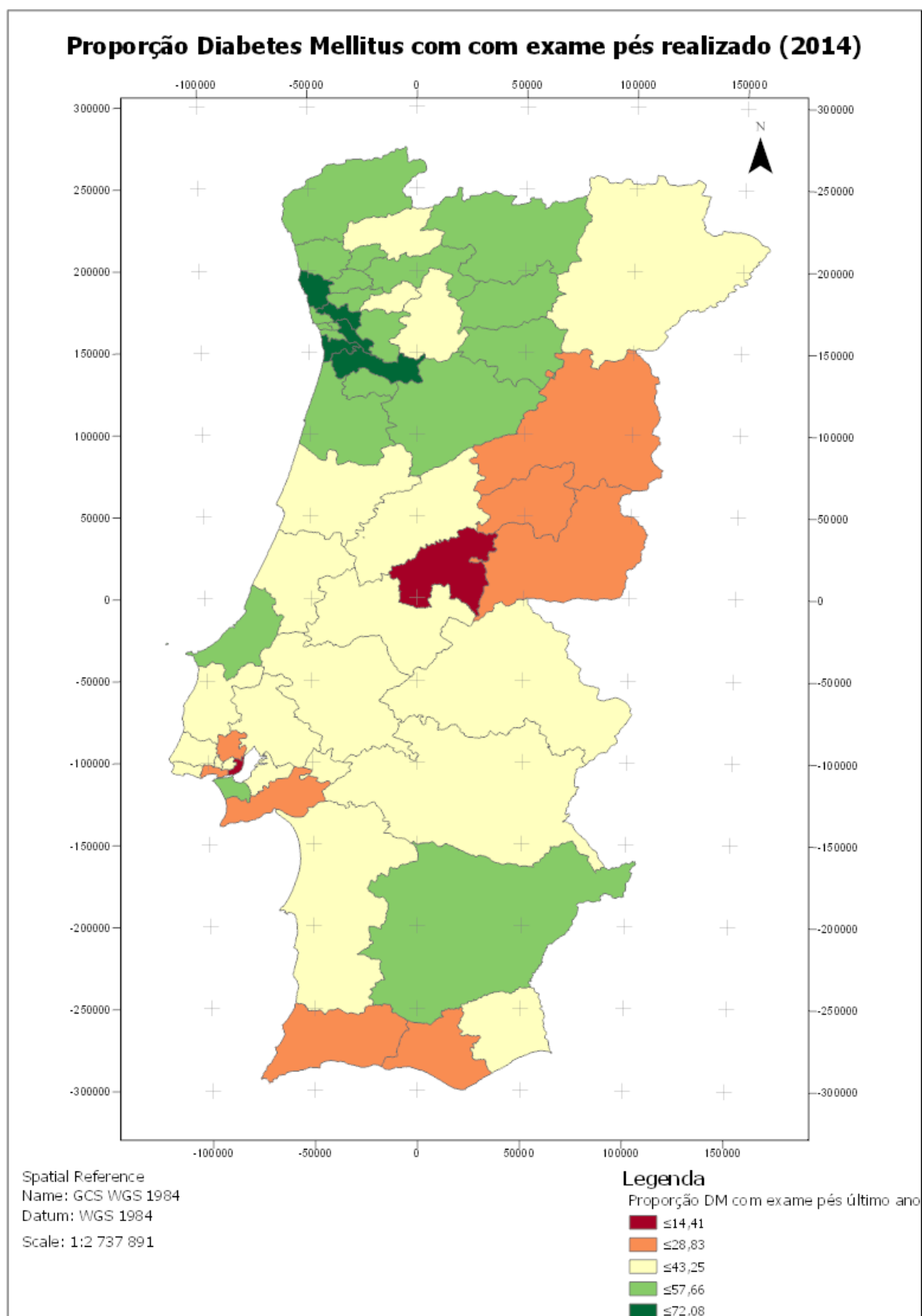
## Anexo II – Dados relativos à Diabetes

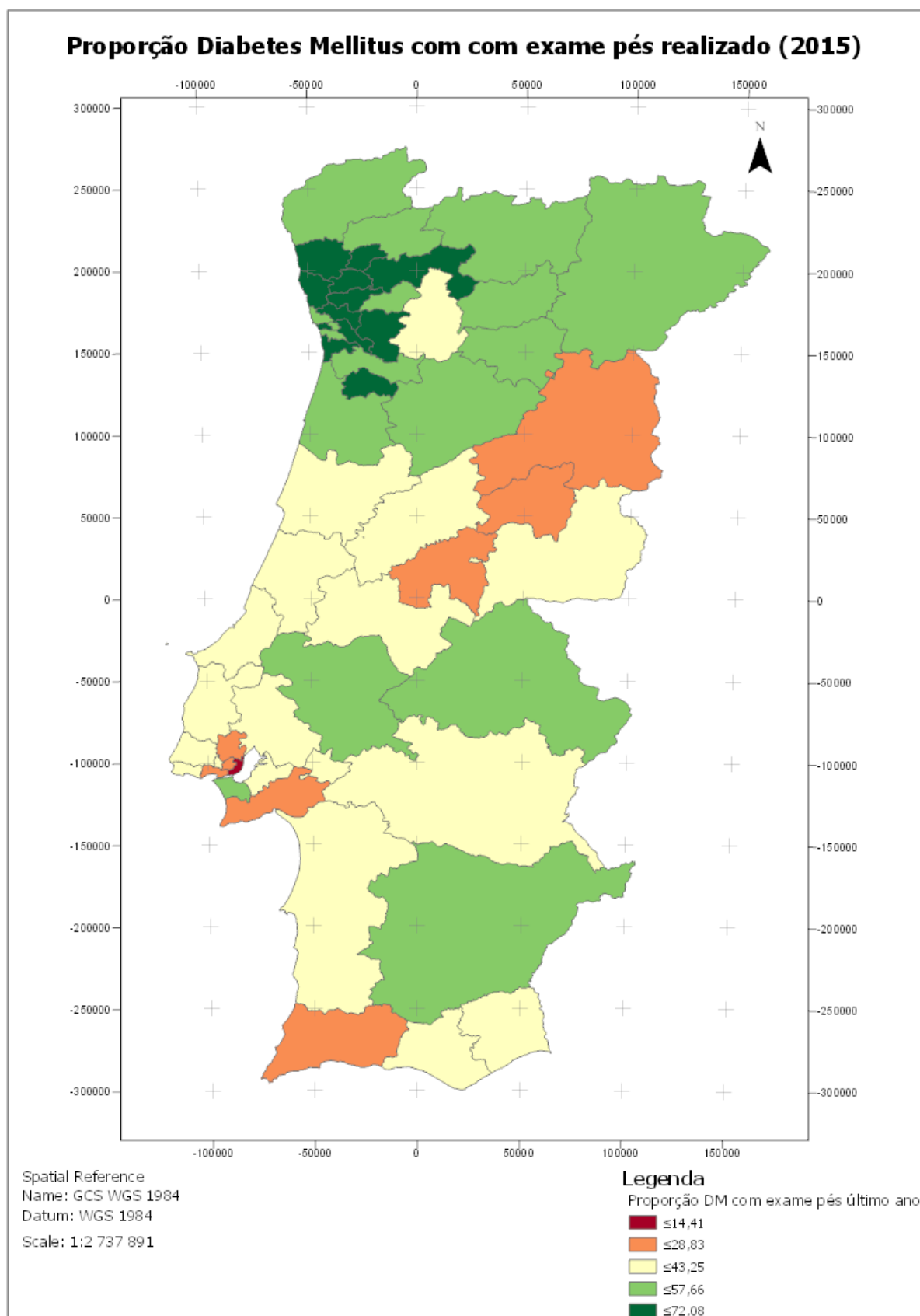


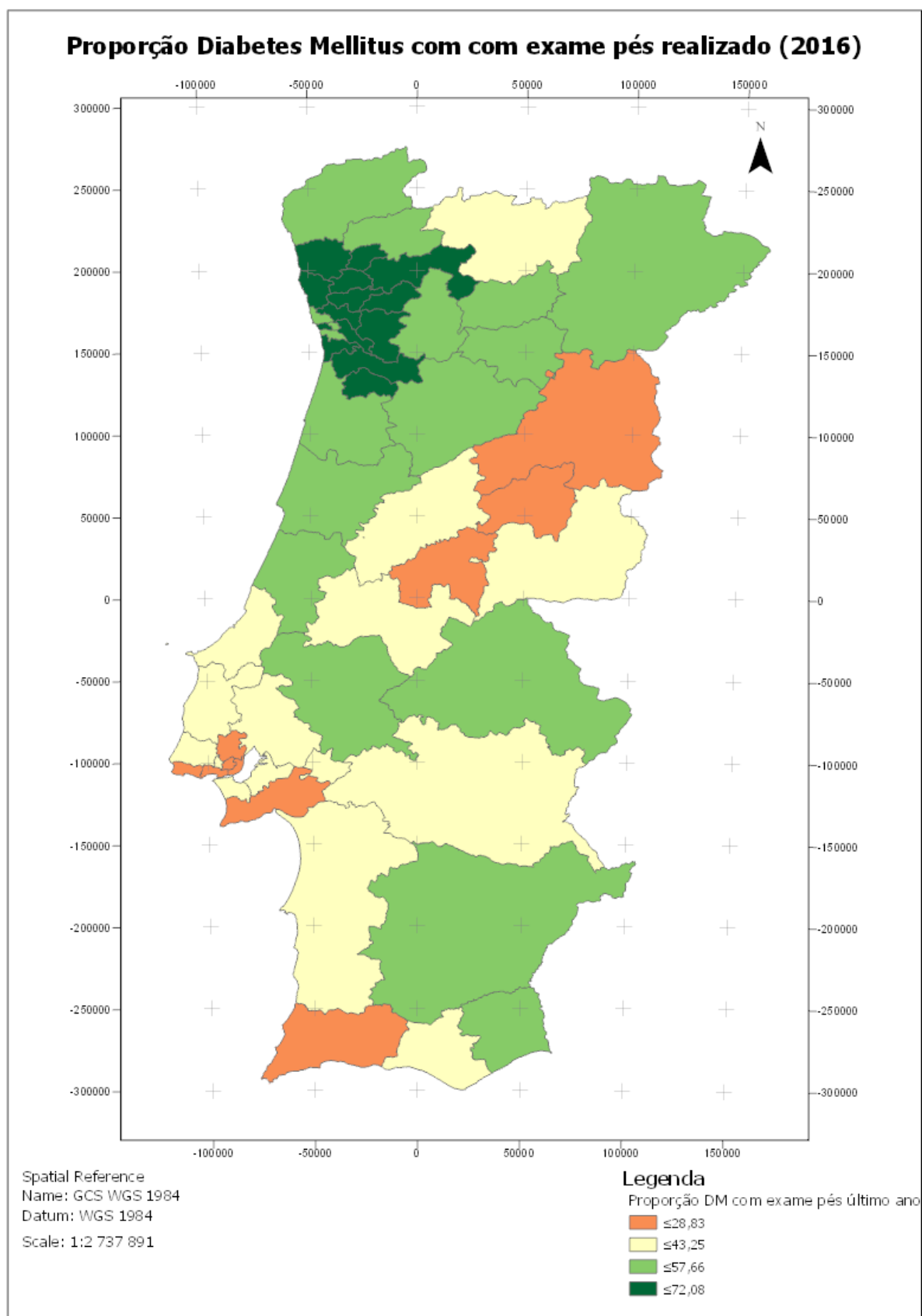




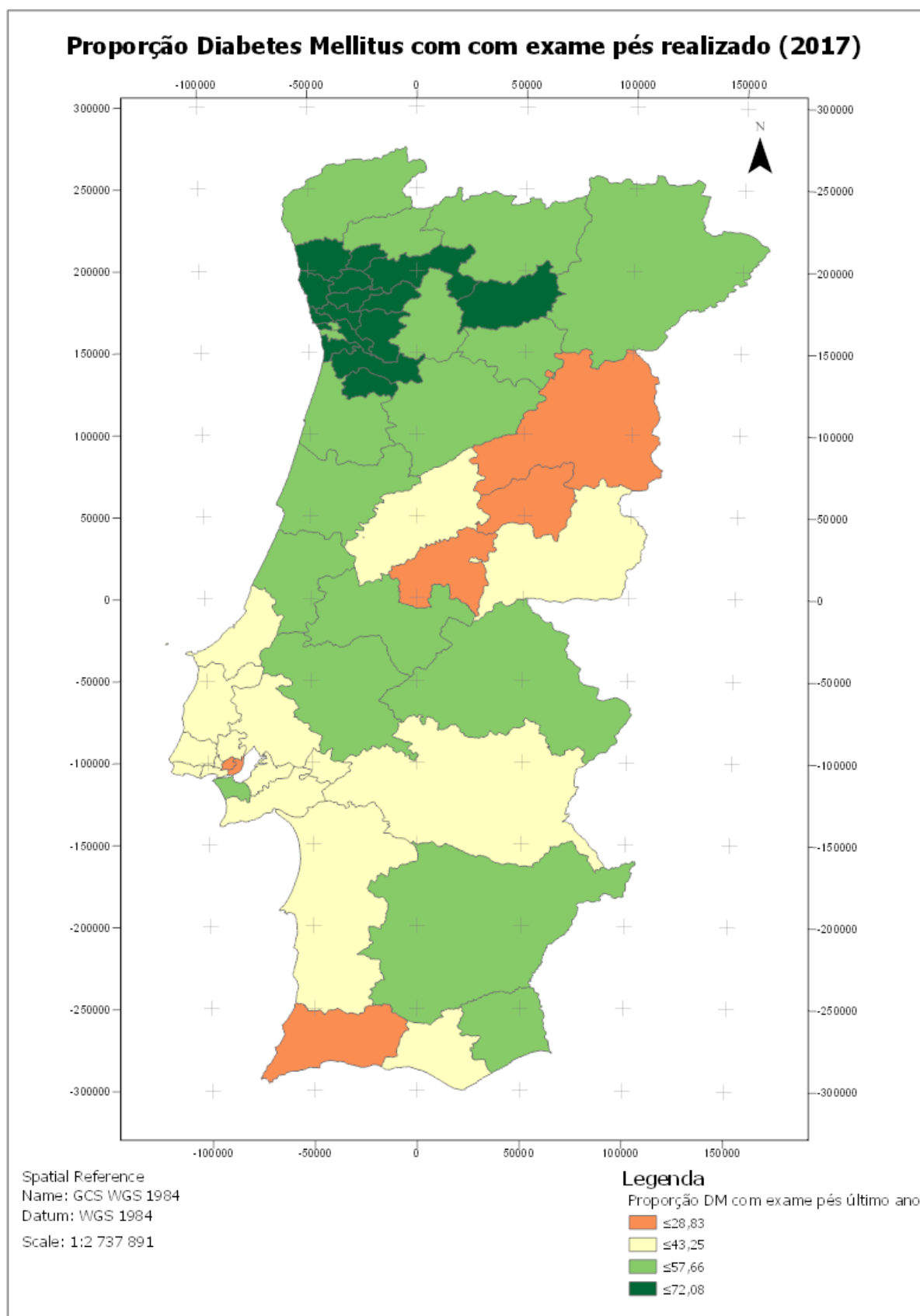


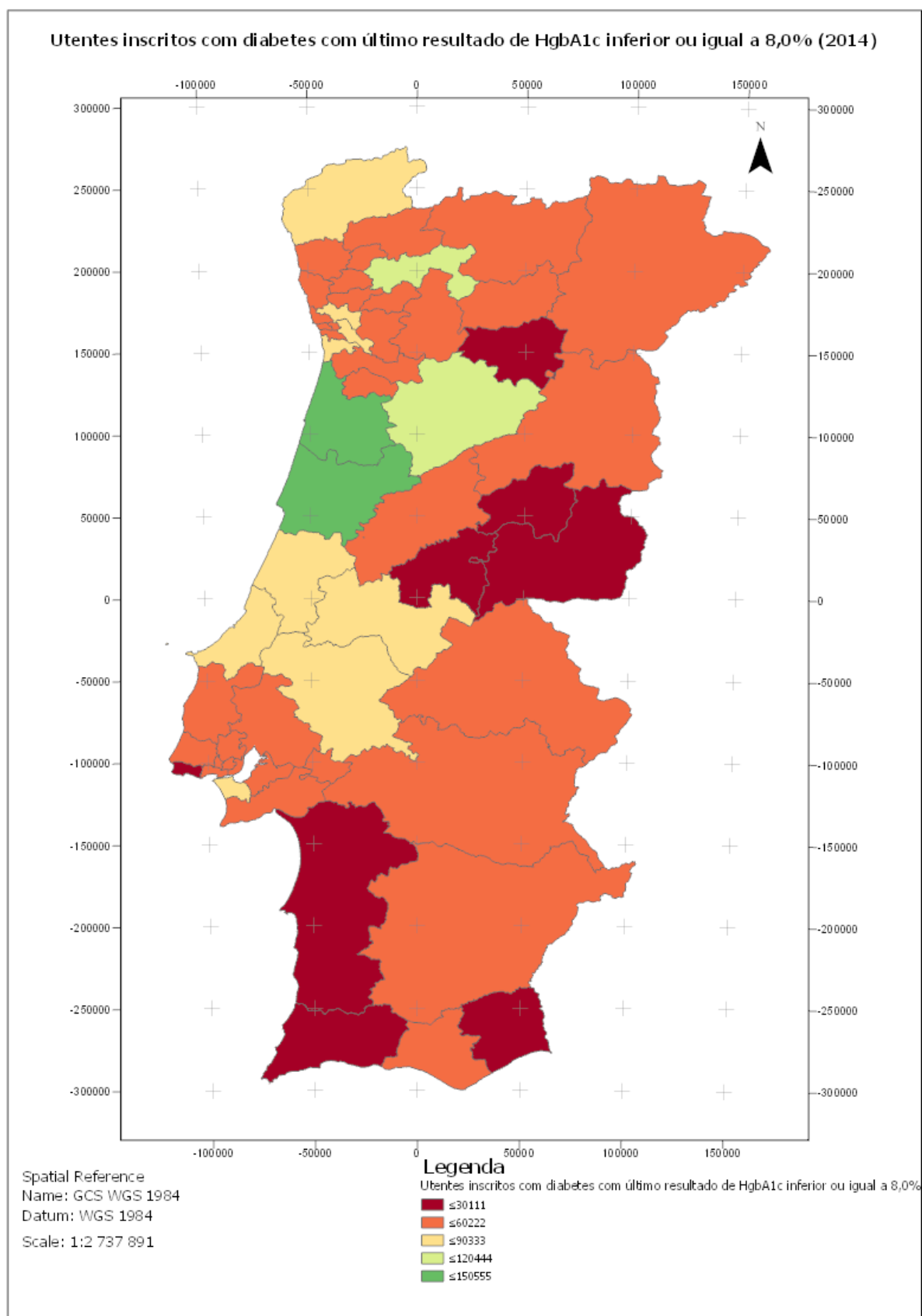


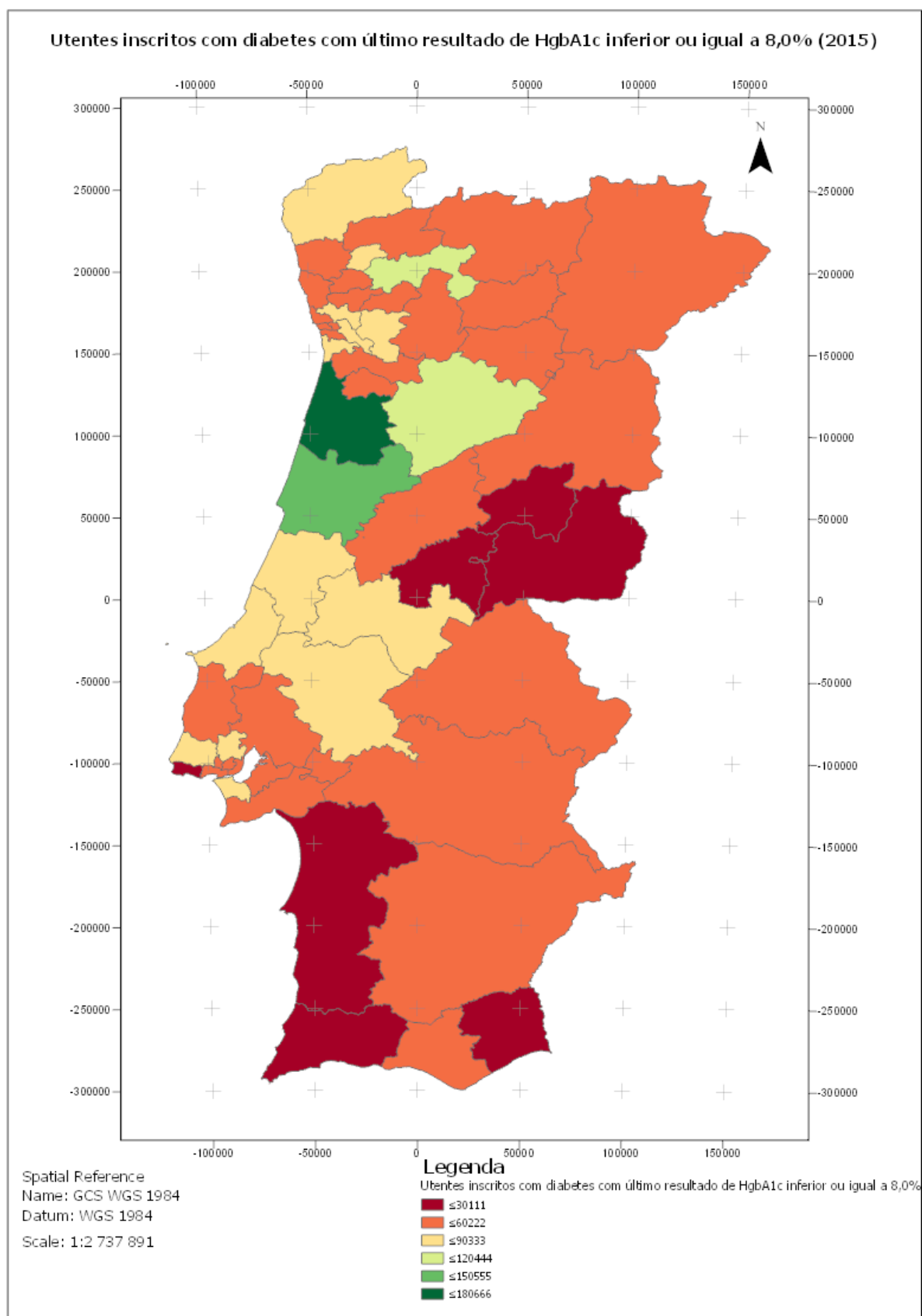


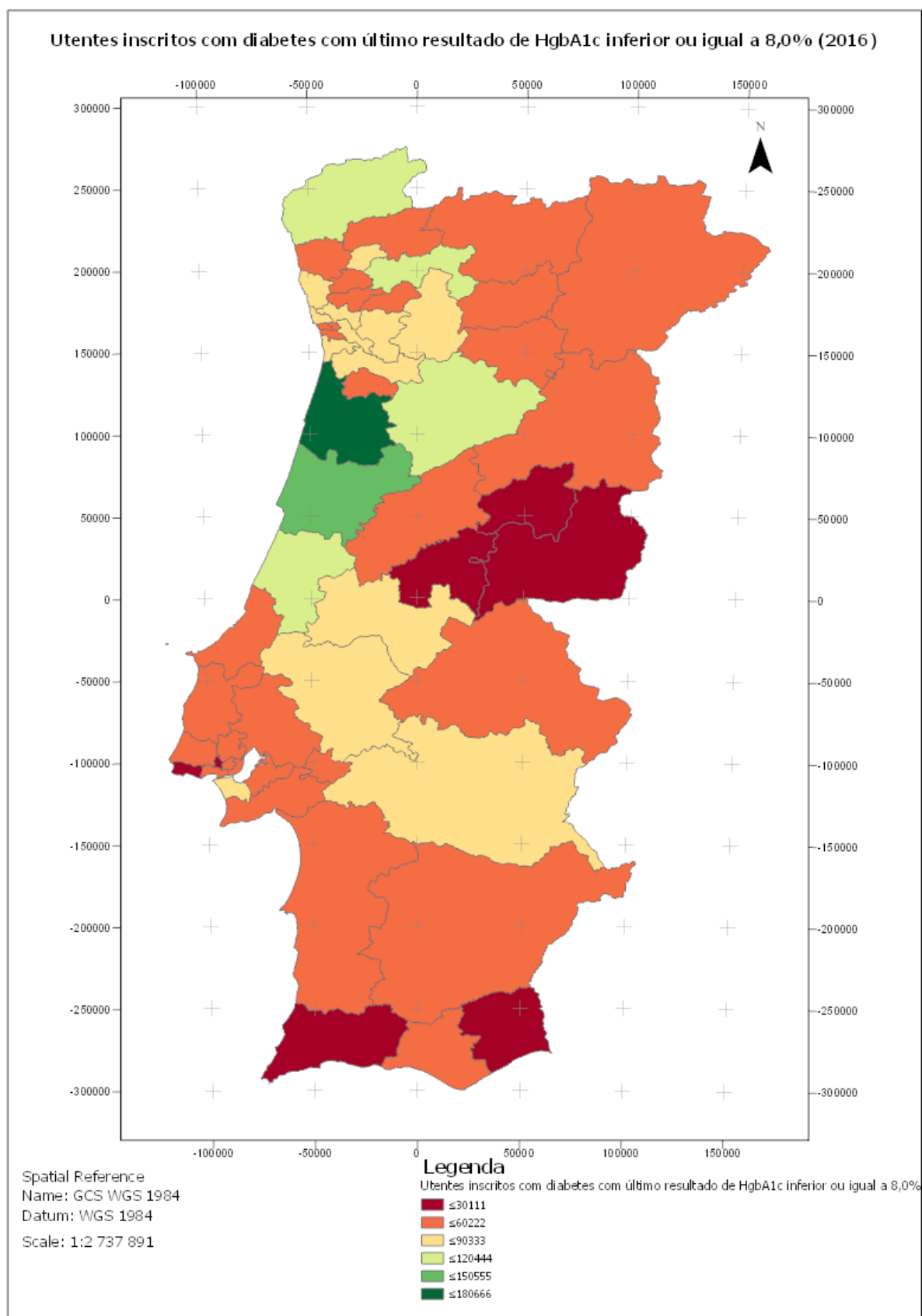


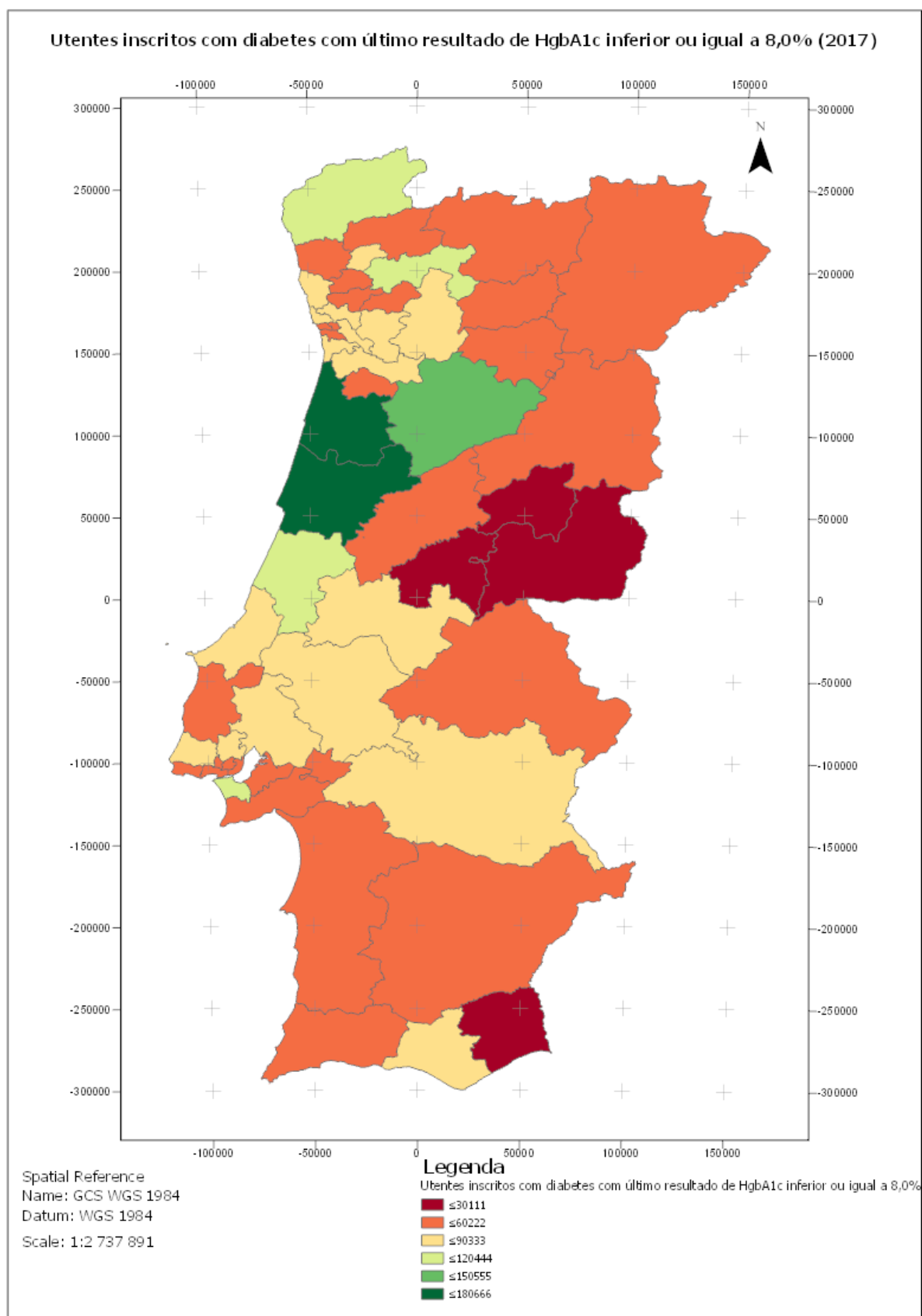


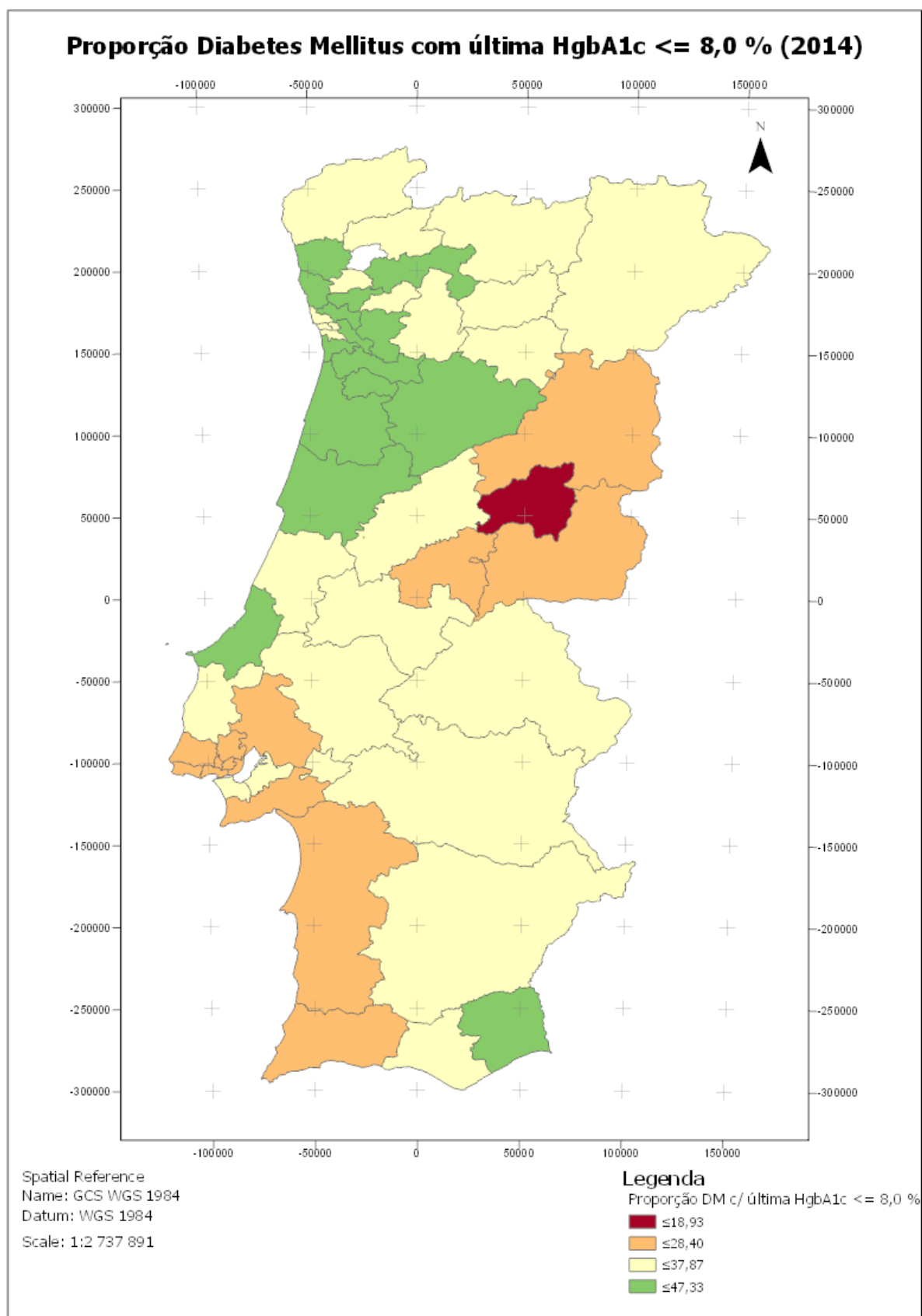


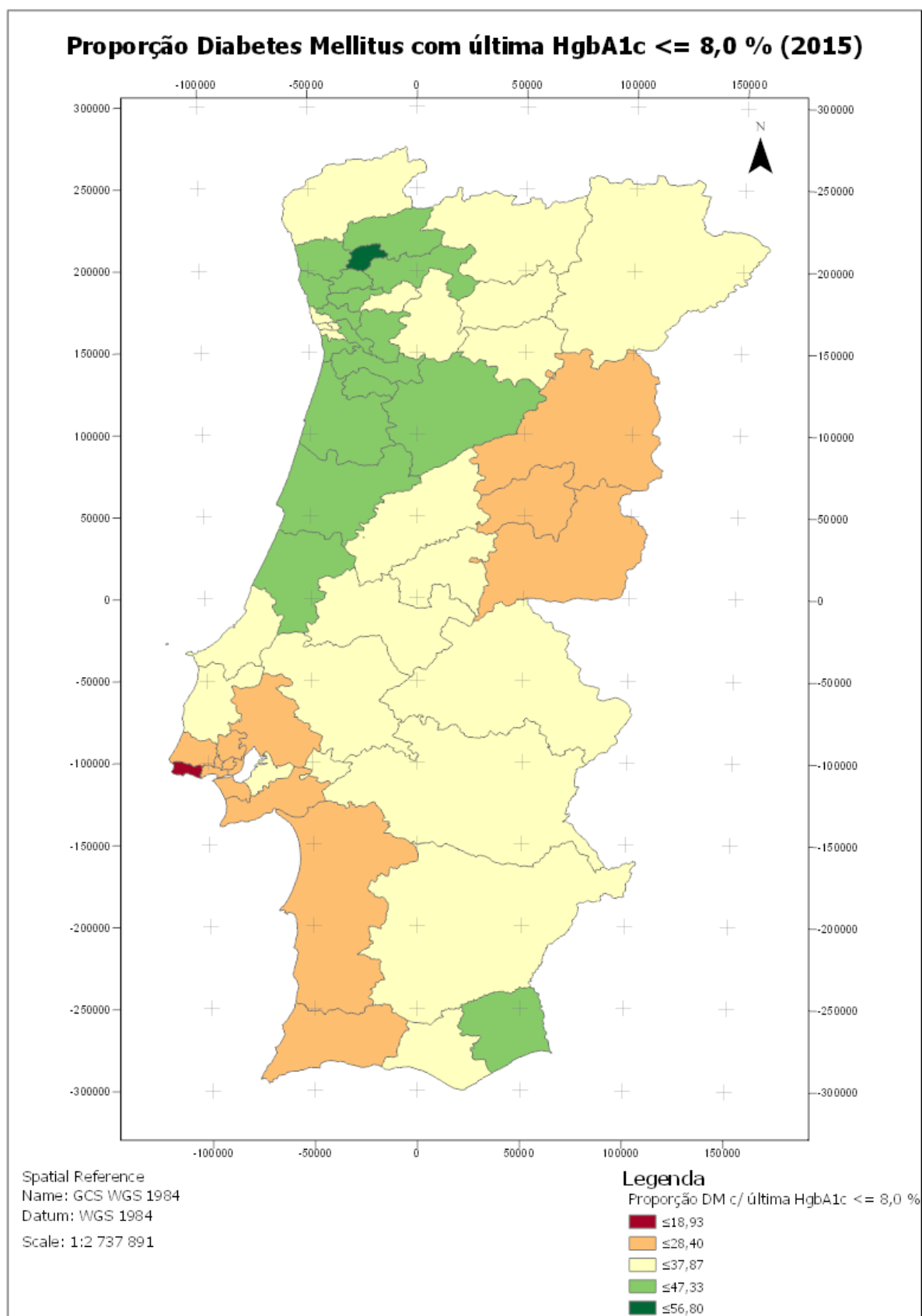




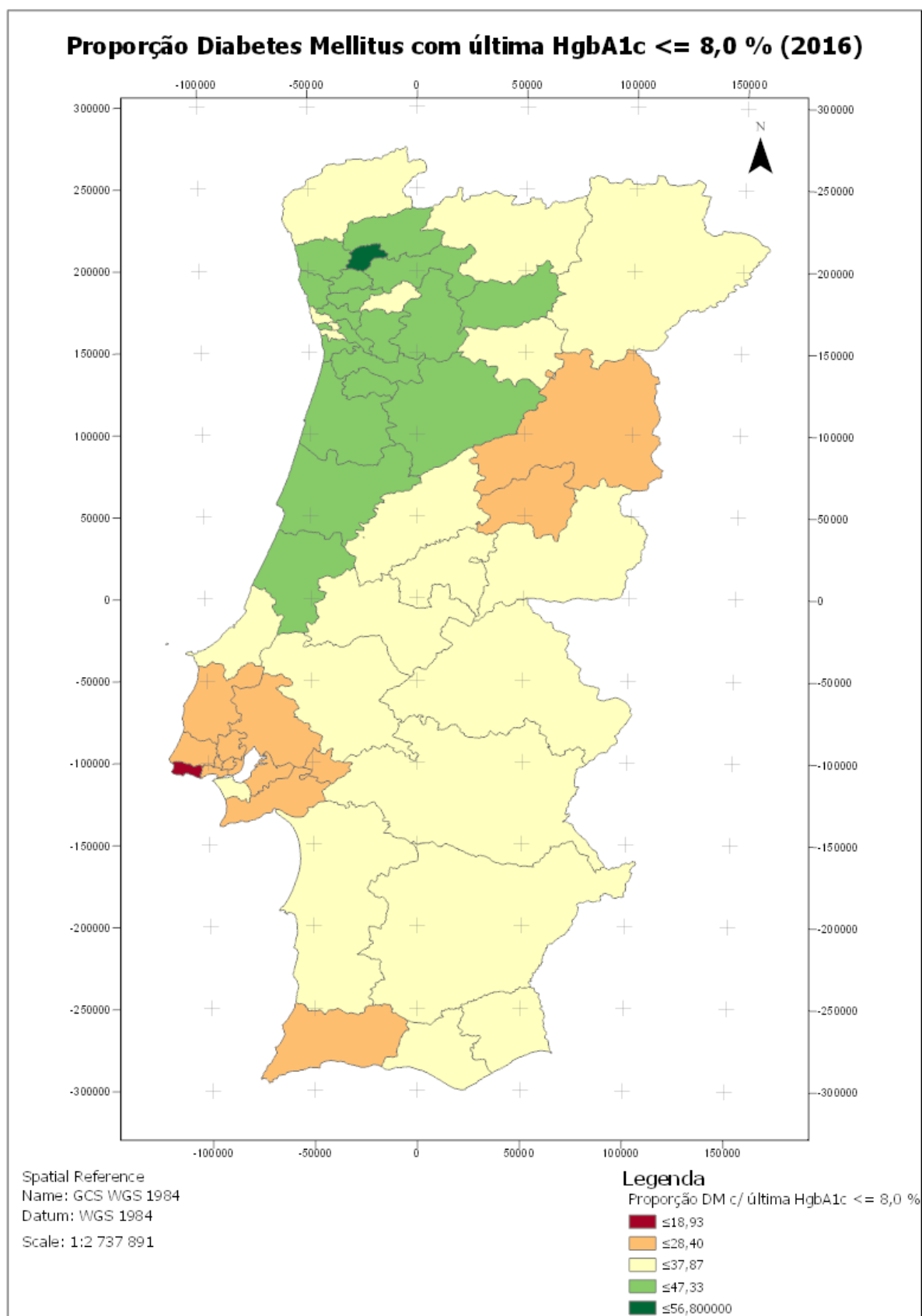


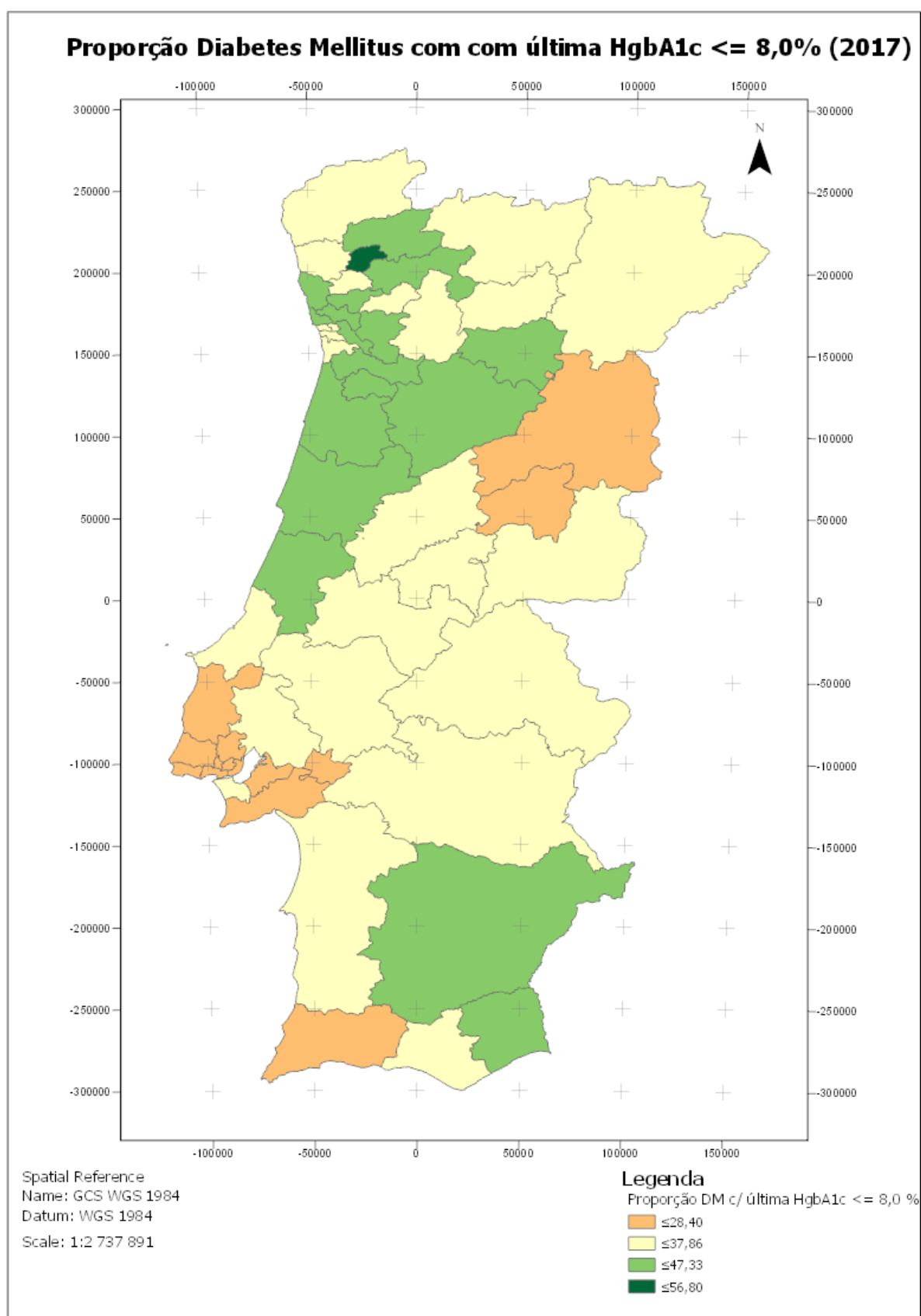




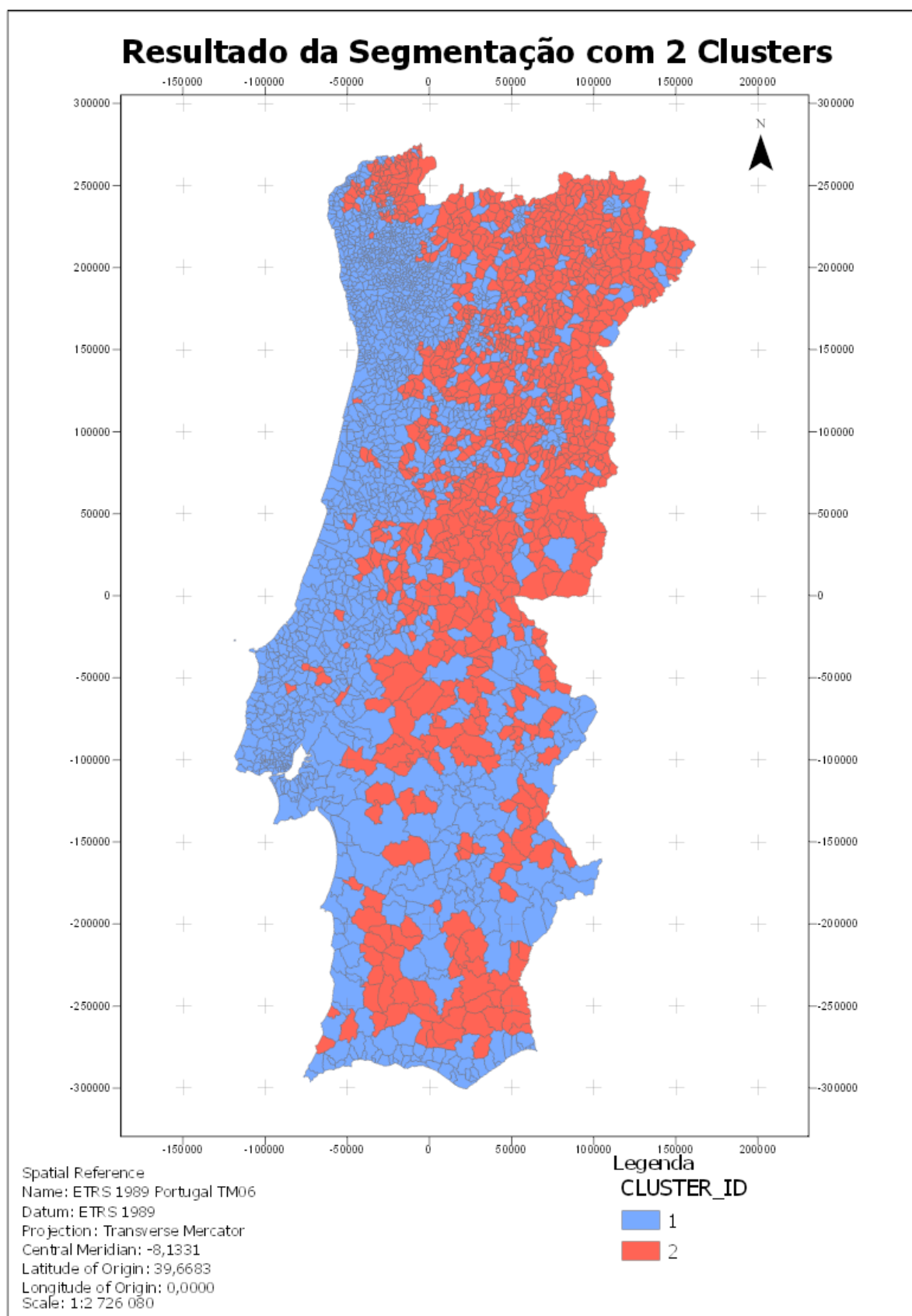


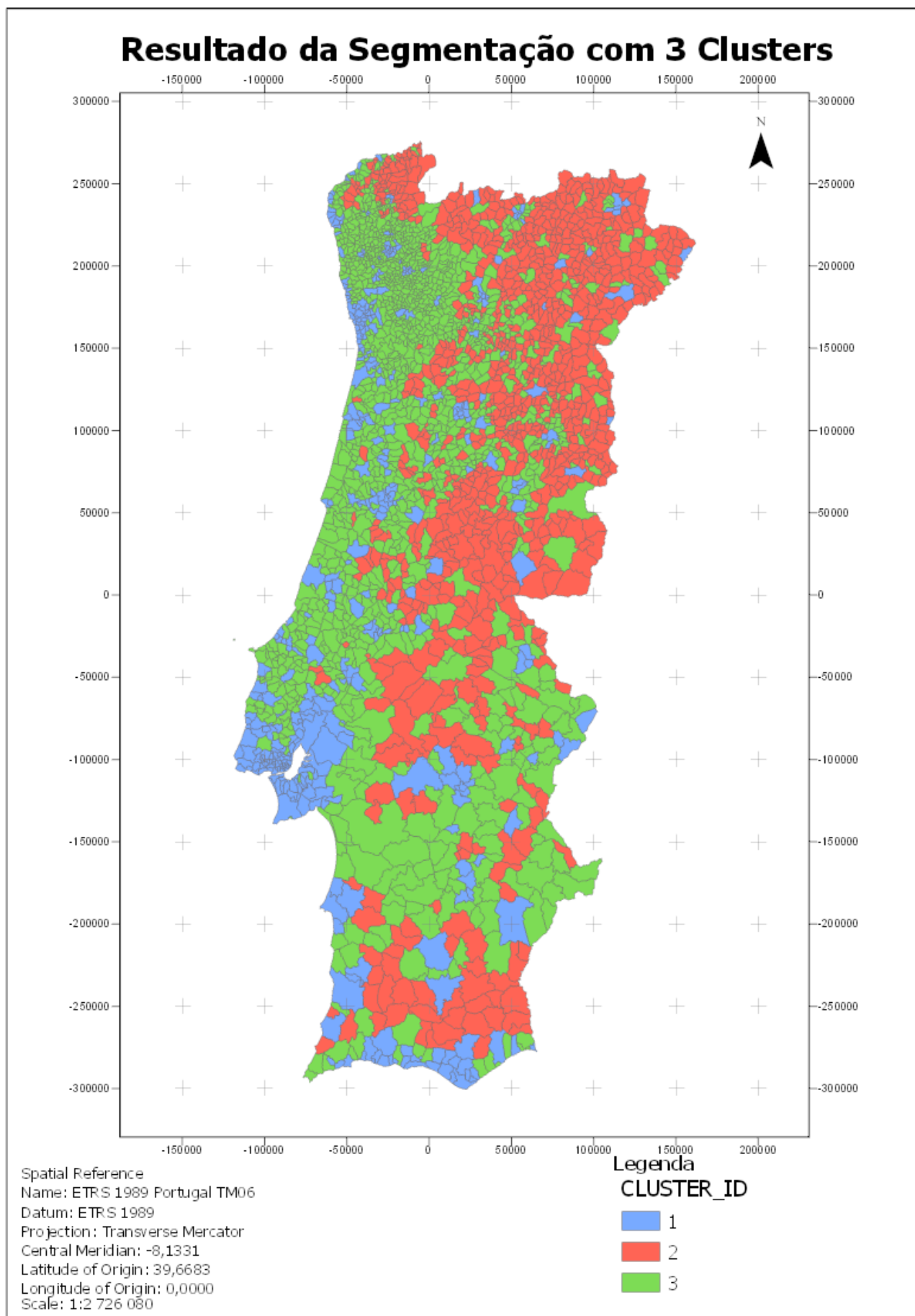






## Anexo III – Resultados da Segmentação com 2 e 3 Clusters





## Anexo IV – Dados utilizados nos Produtos de Informação

*Tabela Anexo IV.1 - Dados provenientes de [11] no formato tabular (Excel) e com resolução espacial ao nível do município*

<b>Dados</b>
Centros de saúde (N.º) por Localização geográfica (NUTS - 2013) e Tipo de serviço (2012)
Enfermeiras/os (N.º) por Local de trabalho (NUTS - 2013) e Sexo (2017)
Médicas/os (N.º) por Local de residência (NUTS - 2013), Sexo e Tipo de médico (2017)
Farmacêuticas/os (N.º) por Local de trabalho (NUTS - 2013) e Sexo (2017)
Farmácias e postos farmacêuticos móveis (N.º) por Localização geográfica (NUTS - 2013) e Tipo de unidade local de farmácia (2017)
Hospitais (N.º) por Localização geográfica (NUTS - 2013) e Natureza institucional (2017)
Óbitos (N.º) por Local de residência (NUTS - 2013), Sexo e Idade (2017)
Atendimentos em serviço de urgência (N.º) nos hospitais por Localização geográfica (NUTS - 2013) (2016)
Atendimentos no serviço de urgência básica - SUB ou no serviço de atendimento permanente ou prolongado - SAP (N.º) nos centros de saúde por Localização geográfica (NUTS - 2013) e Destino do utente (2012)
Consultas médicas na unidade de consulta externa (N.º) dos hospitais por Localização geográfica (NUTS - 2013) e Especialidade da consulta (2016)
Pessoal ao serviço (N.º) nos hospitais públicos de acesso universal e hospitais em parceria público-privada por Localização geográfica (NUTS - 2013) e Tipo de pessoal ao serviço (2016)
Pessoal ao serviço (N.º) nos centros de saúde por Localização geográfica (NUTS - 2013) e Tipo de pessoal ao serviço (2012)
Taxa bruta de mortalidade (‰) por Local de residência (NUTS - 2013) (2017)
Taxa bruta de natalidade (‰) por Local de residência (NUTS - 2013) (2017)
Óbitos (N.º) por Local de residência (NUTS - 2013), Sexo, Grupo etário e Causa de morte (Lista sucinta europeia) (2016)
Partos (N.º) nos hospitais públicos de acesso universal e hospitais em parceria público-privada por Localização geográfica (NUTS - 2013) e Tipo de parto (2016)
Interrupções voluntárias da gravidez legalmente efectuadas (N.º) em estabelecimentos de saúde por Local de residência (NUTS - 2013) (2016)
Quota de mercado do medicamento genérico (PVP - preço de venda ao público - %) por Localização geográfica (2017)

Tabela Anexo IV.2 - Dados do tipo pontual que se encontram no sistema de referência WGS 1984, provenientes de [43]

<b>Dados</b>	<b>Resolução Espacial</b>
Acesso a Consultas Médicas pela População Inscrita (2017)	ACES
Atendimentos em Urgência Hospitalar por Triagem de Manchester (2017)	Instituição Hospitalar
Atendimentos por Tipo de Urgência Hospitalar (2017)	Instituição Hospitalar
Consultas Médicas Hospitalares (2017)	Instituição Hospitalar
Consultas Médicas nos Cuidados de Saúde Primários (2017)	ACES
Taxa de Ocupação Hospitalar (2017)	Instituição Hospitalar
Utentes Inscritos em Cuidados de Saúde Primários (2017)	ACES
Atividade do Programa de Diabetes (2017)	ACES
Atividade do Programa de Hipertensão (2017)	ACES
Administração de Vacinas nos Cuidados de Saúde Primários (2017)	ACES
Atividade do Programa Nacional de Vacinação	ACES

## Anexo V – Dados utilizados no Dashboard dos ACES

Tabela Anexo V.3 - Dados utilizados na criação do Dashboard de Monitorização dos Cuidados de Saúde a nível de ARS e ACES

Dados	Fonte	Breve Descrição
<b>Acesso a Consultas Médicas pela População Inscrita</b>	Transparência	Taxas de utilização e quantidades de acesso a consultas médicas pela população inscrita, apresentando a proporção de utentes inscritos que tiveram consulta médica nos cuidados de saúde primários, no período em análise
<b>Atendimentos em Urgência Hospitalar por Triagem de Manchester</b>	Transparência	Evolução do número de atendimentos em urgência hospitalar, de acordo com a atribuição de prioridade no atendimento aferida pela cor (Triagem de Manchester)
<b>Atendimentos por Tipo de Urgência Hospitalar</b>	Transparência	Monitorização do número de atendimentos em urgência, detalhado por tipo de urgência
<b>Consultas Médicas Hospitalares</b>	Transparência	Monitorização do número de consultas hospitalares, tendo em conta primeiras e subsequentes consultas
<b>Consultas Médicas nos Cuidados de Saúde Primários</b>	Transparência	Evolução de consultas diretas (presenciais) e indiretas (não presenciais)
<b>Taxa de Ocupação Hospitalar</b>	Transparência	Relação percentual entre o total de dias de internamento no ano e a capacidade do estabelecimento.
<b>Utentes Inscritos em Cuidados de Saúde Primários</b>	Transparência	Monitoriza a informação dos utentes inscritos em unidades funcionais
<b>Atividade do Programa de Diabetes</b>	Transparência	Monitorização do programa de acompanhamento de utentes com Diabetes
<b>Atividade do Programa de Hipertensão</b>	Transparência	Monitorização do programa de acompanhamento dos utentes com hipertensão arterial
<b>Administração de Vacinas nos Cuidados de Saúde Primários</b>	Transparência	Número de inoculações de vacinas administradas nos Cuidados de Saúde Primários
<b>Atividade do Programa Nacional de Vacinação</b>	Transparência	Monitorização do cumprimento do Programa Nacional de Vacinação (PNV) e o programa de saúde de adultos Vacina da Gripe
<b>Carta Administrativa de Portugal</b>	DGT	Áreas Administrativas de Portugal



BGRI	INE	Dados sobre os censos efetuados em 2011
------	-----	---

## Anexo VI - Código Python

try:

```

listaDicofres = []

with arcpy.da.SearchCursor("Freguesias_Lisboa",("Freguesias_Dicofre","Shape_Area")) as cursorFreguesias:

    sorteadoFreguesias = sorted(cursorFreguesias)

    tamanho = len(sorteadoFreguesias)

with arcpy.da.SearchCursor("Hospitais_Publicos",("OBJECTID","Freguesias_Dicofre")) as cursorHospitais:

    sorteadoHospitais = sorted(cursorHospitais)

with arcpy.da.SearchCursor("Intersect",("Freguesias_Dicofre","Shape_Area",
"FacilityID","ToBreak","INF_NOME" )) as cursorAreas:

    sorteadoAreas = sorted(cursorAreas)

    for row in sorteadoAreas:

        i=0

        while i < tamanho :

            if row[0] == sorteadoFreguesias[i][0]:

                if row[1] > 0.2 * sorteadoFreguesias[i][1]:

                    listaDicofres.append(row[0])

                i = i + 1

ID_Facility = sorteadoAreas[0][2]

Nome_Facility = sorteadoAreas[0][4]

Tempo = sorteadoAreas[0][3]

Nome_Ficheiro = "Traffic" + "_" + str(sorteadoAreas[0][2]) + "_" + str(int(sorteadoAreas[0][3]))

arcpy.CopyFeatures_management("Freguesias_Lisboa", str(Nome_Ficheiro))

arcpy.AddField_management(Nome_Ficheiro,"Facility_ID", "SHORT")

arcpy.AddField_management(Nome_Ficheiro,"INF_NOME", "TEXT")

arcpy.AddField_management(Nome_Ficheiro,"ToBreak", "SHORT")

with arcpy.da.UpdateCursor(Nome_Ficheiro,("Freguesias_Dicofre","Shape_Area","Facility_ID",
"INF_NOME", "ToBreak")) as cursorAreasFinais:

    if listaDicofres == [] :
```

```
for row in sorteadoHospitais:

    if row[0] == ID_Facility :

        Dicofre_Hospital = row[1]

for row in cursorAreasFinais:

    row[2] = ID_Facility

    row[3] = Nome_Facility

    row[4] = Tempo

    cursorAreasFinais.updateRow(row)

    if row[0] != Dicofre_Hospital:

        cursorAreasFinais.deleteRow()

else:

    for row in cursorAreasFinais:

        row[2] = ID_Facility

        row[3] = Nome_Facility

        row[4] = Tempo

        cursorAreasFinais.updateRow(row)

        if row[0] not in listaDicofres:

            cursorAreasFinais.deleteRow()

del cursorHospitais

except Exception as e:

    print(e.message)
```